DOCUMENT RESUME

ED 161 756 SE 025 248

TITLE	Case Studies in Applied Mathematics.
INSTITUTION	Mathematical Association of America, Washington, D.C.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	76
NOTE	438p.; Pages 326-343 removed due to copyright restrictions; Not available in hard copy due to marginal legibility of original document
AVAILABLE FROM	The Mathematical Association of America, 1529 Eighteenth St., N.W., Washington, D.C. 20036 (no price quoted)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Behavior Patterns; *College Mathematics; Communicable Diseases; Computers; Ecology; Heat; Higher Education; *Instruction; *Mathematical Applications; *Mathematical Models; Political Power; Population Trends; Power Mechanics; Statistics; *Teaching Guides
IDENTIFIERS	*Committee on the Undergraduate Program in Math

ABSTRACT

This collection of nine case studies in applied mathematics was written primarily for the use of the instructor by a Conference sponsored by the Committee on the Undergraduate Program in Mathematics (CUPM). Each chapter contains exercises of varying degrees of difficulty and several include student projects. The materials were used on a trial basis and the results of these experiences are reported. The first chapter discusses the process of applied mathematics. The case studies are: (1) measuring power in weighted voting systems; (2) a model for municipal street-sweeping operations; (3) a mathematical model of renewable resources; (4) some examples of mathematical models for the dynamics of several-species ecosystems; (5) population mathematics; (6) MacDonald's work on Helminth Infections; (7) modeling linear systems by frequency response methods; (8) network analysis of steam generator flow; and (9) heat transfer in frozen soil. (MP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161756

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Leonard Gillman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

CASE STUDIES IN APPLIED MATHEMATICS

COMMITTEE ON THE UNDERGRADUATE PROGRAM IN MATHEMATICS

MATHEMATICAL ASSOCIATION OF AMERICA

1976

SE 025 248

CASE STUDIES IN APPLIED MATHEMATICS

Supported by
NATIONAL SCIENCE FOUNDATION

Committee on the Undergraduate Program in Mathematics

1976

COMMITTEE ON THE UNDERGRADUATE PROGRAM IN MATHEMATICS

William F. Lucas
Cornell University
Chairman

Donald Bushaw
Washington State University

Arthur P. Mattuck
Massachusetts Institute of Technology

Juanita J. Peterson
Laney College

T. A. Porsching
University of Pittsburgh

F. S. Roberts
Rutgers, The State University

Alex Rosenberg
Cornell University

Maynard D. Thompson
Indiana University

Alan C. Tucker
State University of New York at Stony Brook

June P. Wood
University of Houston

Gail S. Young
University of Rochester

Henry O. Pollak
Bell Telephone Laboratories, Inc.
ex officio

TABLE OF CONTENTS

Preface 1

Chapter 1 The Process of Applied Mathematics
Maynard D. Thompson 4

Chapter 2 The Trial Teaching of the Modules 22

Chapter 3 Measuring Power in Weighted Voting Systems
William F. Lucas 42

Chapter 4 A Model for Municipal Street Sweeping Operations
Alan C. Tucker and L. Bodin 107

Chapter 5 A Mathematical Model of Renewable Resource Conservation
Colin C. Clark 151

Chapter 6 Some Examples of Mathematical Models for the Dynamics of Several-Species Ecosystems
H. R. van der Vaart 186

Chapter 7 Population Mathematics
Frank C. Hoppensteadt 293

Chapter 8 MacDonald's Work on Helminth Infections
Donald Ludwig and Benjamin D. Haytock 313

Chapter 9 Modeling Linear Systems by Frequency Response Methods
William F. Powers 344

Chapter 10 Network Analysis of Steam Generator Flow
Thomas A. Porsching 372

Chapter 11 Heat Transfer in Frozen Soil
Gunter Meyer 410

PREFACE

The utility of mathematics as an aid to solving problems that arise in the physical sciences is widely appreciated. Almost as widely recognized is the stimulation that mathematics has received from attempts to solve problems that arise in other fields. Especially in the last decade, this interplay between mathematics and the physical sciences has extended with increasing vigor into the biological, behavioral and social sciences, and even into some areas of the humanities.

Until quite recently undergraduate instruction in the applications of mathematics has tended to focus on the development of mathematical techniques that have proved useful in studying problems arising in the physical sciences. Often the problems themselves are quite idealized and nearly all attention is devoted to studying questions that are initially posed in a mathematical form. An alternative which would provide undergraduate students an experience corresponding more nearly to the use of mathematics outside of the undergraduate classroom would consist of studying realistic problems--even though this may mean that the problems cannot be solved completely and neatly in the usual sense--and of carrying the study from beginning to end, i.e., through all the phases of the process of applying mathematics.

Support for proposals of the latter sort has been provided by individuals and professional organizations for several years. However, implementation of these recommendations has proved to be less tractable than had been anticipated. For example, at a Conference sponsored by the Committee on the Undergraduate Program in Mathematics (CUPM) in 1972 and attended by mathematicians involved in undergraduate instruction there was widespread agreement with the proposals, but skepticism that implementation could proceed without appropriate materials. It was decided, therefore, to undertake a project to prepare sample materials for use by college faculty in undergraduate courses.

Concurrent with these considerations of the flavor and content of undergraduate applied mathematics courses, there have been discussions of the extent to which open-ended model building by students is profitable in the undergraduate classroom. The concept of the project was expanded to include trial testing and evaluation, and, in particular, some experimentation with open-ended use of the materials. The results of the testing are discussed in some detail in Chapter 2 "Trial Teaching of the Modules."

The materials were written primarily for the use of the instructor. Some of the authors first encountered the problems about which they write in the setting of industrial research and development. Each of the problems has attracted attention in the scientific literature other than in mathematics books and journals. Obviously many other topics would be equally appropriate, and the present collection reflects to some extent the preferences of the authors and the editor.

Each chapter contains exercises of varying degrees of difficulty, and several include projects which require more time, energy and creativity on the student's part. The references at the end of each chapter serve as an entry into the (substantial) literature in each area.

Chapter 3, "Measuring Power in Weighted Voting Systems," illustrates the use of numerical indices in estimating and comparing the power of units of a political system using a weighted voting scheme in which the effective influence of voters varies from one voter to another. The early sections use only elementary set theory and combinatorics. The later sections require more mathematical sophistication, and in the section on Computational Aids, a knowledge of computing. This chapter contains many projects and an extensive bibliography.

Chapter 4, "A Model for Municipal Street Sweeping Operations," provides an example of the use of mathematical modeling to improve the efficiency of urban services. The mathematics used, graph theory, has gained increasing recognition as a valuable tool in the study of social and urban problems. This chapter illustrates well the need for effective computational algorithms: The model is only useful if one has a means of obtaining the best routes with reasonable efficiency. Section 10 contains a variety of exercises. Another reference to applications of this type is Discrete Mathematical Models, with Applications to Social, Biological and Environmental Problems by Fred S. Roberts (Prentice-Hall, Inc., 1976).

Chapter 5, "A Mathematical Model for Renewable Resource Conservation," is concerned with situations which have both economic and biological aspects, and the basic question considered is one of determining the optimal use of renewable natural resources. The mathematical prerequisite is elementary calculus; the relevant economic and biological concepts are developed within the chapter. Each section contains exercises and there are notes and references at the end of the chapter.

Chapter 6, "Dynamics of Several Species Ecosystems," surveys a topic with a rich history in both the mathematical and biological literature. Calculus and differential equations play a major role in the chapter. There is a self-contained treatment of discrete time models using difference equations in Section 1.6. Questions for discussion are scattered throughout the chapter, and there are some hints and remarks regarding the questions in Appendix C. Appendix A introduces the technique of linearization and related mathematical ideas. Some of the more recent textbook and monograph literature is surveyed in the Introduction, and Appendix B includes a more comprehensive annotated bibliography. One of the references, d'Ancona, The Struggle for Existence, is almost essential for reading the chapter. The Preface and Introduction to the chapter give a more detailed overview and a useful historical perspective.

In Chapter 7, "Population Mathematics," a model for the evolution of a single species population is formulated. It is supposed that age-specific fertility and mortality rates are known, and the stable age distribution of the population and population waves are discussed. A knowledge of elementary differential equations and Laplace transforms is assumed. There are exercises and projects, some of which require a knowledge of elementary numerical analysis. The book Introduction to the Mathematics of Population by N. Keyfitz is a very useful reference.

Chapter 8, "MacDonald's Work on Helminth Infections," gives an example of the mathematical modeling activities regularly carried out by researchers working in other fields. MacDonald's paper originally published in The Transactions of the Royal Society of Tropical Medicine and Hygiene is reproduced. The rest of the chapter indicates some of the ways in which the mathematical ideas introduced in the paper can be recast so as to be more familiar to mathematicians. A first draft of the article was written by Donald Ludwig. Benjamin Haytock was one of those who tested the first draft, and he and Ludwig collaborated on the revised version. Although the explicit mathematical prerequisites are not heavy, elementary differential equations and a little probability, some mathematical sophistication is needed to follow the arguments.

Chapter 9, "Modeling Linear Systems by Frequency Response Methods," illustrates the application of frequency response techniques to human behavior modeling. A knowledge of elementary linear differential equations, including Laplace transform methods is assumed. There is a project (Section 4) involving the analysis of actual data using the methods developed in the chapter.

In Chapter 10, "Network Analysis of Steam Generator Flow," a model of the flow of liquid in a network of tubes in a steam generator is formulated and studied. A simplified version of the model leads to mathematical problems which can be solved numerically. The mathematics used includes network (graph) theory and elementary partial differential equations. There is a prefatory "To the Instructor" section which provides a detailed discussion of references and prerequisites.

In Chapter 11, "Heat Transfer in Frozen Soil," a problem which arose in the construction of the Alaska oil pipeline is studied. A model is constructed from the basic principles of heat conduction, and specific mathematical problems, boundary value problems, are formulated. The early sections of the chapter require only calculus, although a familiarity with elementary differential equations would be useful. Section 5 is considerably more advanced. An algorithm suitable for numerical computation is introduced.

The project will have been successful if the reader finds these examples, or parts of them, useful in his (or her) undergraduate courses. It will be even more successful if the reader is stimulated to take an interesting problem from some other academic area or from the world of government, business, and everyday life and to use it as the basis for a chapter similar to these. We are interested in your experiences in using these materials. Please send your comments to: MAA Special Projects Office, Department of Mathematics, California State University, Hayward, California 94542.

This project could not have been carried out without the dedication and cooperation of the authors whose articles comprise the major part of the book. The faculty members who taught from the preliminary versions of the materials provided many helpful comments. In addition, at the very early stages, Werner Rheinboldt and Henry O. Pollak provided valuable aid in conceptualizing the project as did Paul Knopp of the CUPM Central Office in organizing and expediting it. Past Chairman of CUPM, Alex Rosenberg and Donald Bushaw, encouraged and supported the project, and Katherine Magann and George Pedrick of the MAA Special Projects Office prepared the manuscripts for publication. To all of the above, and to many others who contributed in various ways, our sincere thanks. The financial support of the National Science Foundation is gratefully acknowledged.

Maynard Thompson
Bloomington, Indiana
June, 1976

Chapter 1
THE PROCESS OF APPLIED MATHEMATICS

Maynard Thompson
Indiana University

1. Introduction

In discussing the applications of mathematics in the undergraduate curriculum one is inclined to begin with a precise definition of what is meant by applied mathematics. There are many informative discussions of the nature of what is commonly referred to as applied mathematics, for example [4], [10], [12] and [13] listed in the references, and it is not our purpose here to expand on these. Instead, we will initially adopt a relatively informal approach, avoiding definitions and relying on comments and examples to convey the ideas. It is our hope that the reader will acquire from this chapter an appreciation of the process which is exemplified in the modules which comprise the major part of this book.

If one views mathematics in a very broad sense as rigorous, logical thinking, then the scholar or student whose approach to his discipline involves a precise identification of the concepts, definitions and assumptions, and the deduction of consequences from these assumptions through logical argument is doing applied mathematics. Our point of view here is much narrower. We have in mind only those studies in which a situation arising outside of mathematics is clarified by making appropriate use of mathematical ideas and techniques which are accessible to an undergraduate.

When we speak of the applications of mathematics it is clear that we have in mind applying mathematics to something. That something could, of course, be another branch of mathematics. However, we will exclude from our discussion the use of the mathematics originating in one branch (e.g., probability) to the study of problems arising in another (e.g., analysis). This choice reflects the purpose of these materials and does not imply that such uses are unimportant or uninteresting. Rather, they are simply outside of the scope of this collection. We concentrate on the use of mathematics to study situations which arise in the everyday world of business, government, industry, etc., or in another discipline. There is an expanding range of situations for which mathematics yields useful conclusions, and even more which have been studied using mathematics. The goal of the process of applying mathematics is to learn something about the situation which either was unknown or not firmly supported prior to the mathematical investigation.

To be more precise, we ought to speak of the use of the ideas and techniques of the mathematical sciences rather than simply those of mathematics, which for many means pure mathematics. Indeed, it may be that statistics and computer simulation are the most appropriate tools to use on a problem arising in urban planning. Many of the more interesting problems are sufficiently complex that one needs to utilize concepts from several branches of the mathematical sciences to make progress. This amalgamation of methods from several mathematical areas is one of the common characteristics of applied mathematics. In the future when the phrase "use of mathematics" occurs, it is to be interpreted in the broad sense just described.

The application of mathematical methods to a precisely formulated problem is one part of the process of applying mathematics. Another important component of that process has come to be referred to as modeling. Although the term has been used in many different ways in the literature, it is a useful descriptor and we will use it freely. Our use of the term will become precise in the next section.

2. Applied Mathematics as a Process

Applied mathematics is a sciences which is concerned with the interaction between pure mathematics and another subject--another academic discipline or some aspect of the everyday world. One of the more useful analyses of the way in which this interaction takes place is due to Dr. H. O. Pollak. (Another, ascribed to J. L. Synge, may be found in [4] p. 12.) This organization provides a profitable starting point for our discussion, and it is appropriate to summarize it here. Briefly, the process can be viewed as consisting of five phases:

I Identification of the problem in the scientific setting.

II Formulation of a mathematical model.

III Solution of the mathematical problems that arise in the study of the model selected.

IV Development of algorithms and associated computer programs for relevant computations.

V Explanation and interpretation of the results in the context of the original problem and the communication of this information to the interested audience. Evaluation of the results.

It is uncommon for a mathematician to play a significant role in Phase I. Usually the situation is being studied by a scientist (or manager or doctor or ...) and it is this individual who recognizes the importance of the issues and the possible relevance of mathematical techniques. It may happen that the subject is studied intensively, perhaps for an extended period, data are collected, and the results summarized in empirical laws before a systematic effort is made to provide supporting theory. In business and industrial settings it is not uncommon for problems to be solved solely on the basis of experience or ad hoc techniques prior to the recognition that mathematical methods may be helpful. Recently, however, the utility of mathematics is more widely appreciated and it is increasingly common to turn quickly to the search for mathematical solutions to problems. Since this is written for mathematicians it is worthwhile to add the cautionary note that the minor role usually played by mathematicians in this phase is very appropriate. Indeed, at present few mathematicians have the scientific knowledge to judge the importance of the issues or to interpret the results of experiments. This may well change as

10

more mathematicians move from the fringes to the center of active research in other disciplines.

Phase II, the formulation of a mathematical model, is frequently the most crucial and most difficult part of the entire process. Usually it is a very creative activity carried out by a mathematically knowledgeable scientist or by a scientist and mathematician working jointly. Model building, or theory construction as it is sometimes called by social scientists, consists of examining the situation carefully, identifying what is important and what is not, and selecting (or creating) a suitable mathematical structure. A model has two components: a mathematical structure (primitive terms, definitions, and axioms), and an identification between the concepts of the real situation and those in the mathematical system. In general the particular structure selected is chosen because it has some theorems (predictions) which are known to be consistent with the data of the original situation. Of course, the goal is to use the mathematical system to deduce new information about the situation or to provide firm support for known results. Generally, in the process of identifying those aspects of the original situation which are to be retained one also simplifies it as much as possible. In this phase such simplifications are made on a scientific and not a mathematical basis. The meaning of each simplification or idealization should be carefully considered with respect to its meaning in the original setting. Obviously it is essential that the simplifications not be so radical that the theorems of the related mathematical system cease to provide valid predictions about the actual situation. On the other hand, it may be necessary that some simplification take place in order that the resulting mathematical system be manageable. The problems of deciding what is important and what is not and which simplifications are legitimate and which are not are major ones and require experience and ingenuity. Clearly the activity must be carried out by an investigator who is thoroughly familiar with the actual situation and the basic scientific principles of the field.

Depending on the situation being investigated there may be several different mathematical structures which provide useful mathematical models. For example, a situation in economics may involve consumer demand as one of the quantities of interest. The model may be significantly different depending upon whether this demand is viewed as a deterministic or probabilistic quantity, and the conclusions based on the model may differ as well. In many cases there will be several alternative models, and there may well be no "best" model. That is, one model may lead to predictions of one sort which agree well with observations while another model leads to predictions of a different sort which agree well with experiments. An example of this from elementary physics is the dual wave and particle models for light. The wave model for light provides explanations for the main phenomena of physical optics: reflection and refraction, dispersion, polarization, and diffraction. However, the photo-electric effect which is difficult to understand in the wave model, is perfectly comprehensible in the particle model.

The creation of a mathematical model includes the selection of a mathematical structure and a correspondence between that structure and the original situation. Thus specific questions regarding the original situation carry over into specific mathematical questions. Phase III is concerned with the study of these mathematical questions. It is this activity that is usually thought of as applied mathematics and which provides the content for most applications oriented undergraduate courses. Actually much of what goes on in this phase is indistinguishable on the surface from pure mathematics, only the motivation is different. It is important, however, to keep in mind that the mathematical problem has a connection with the physical (or social or everyday) world. If the problem must be modified for mathematical reasons, then the relationship of the modified problem with the original one, and consequently with the physical world, must be carefully analyzed.

It is easy to formulate apparently straightforward models for relatively simple situations which lead to extremely difficult mathematical problems. Sometimes these problems fit neatly into a well understood mathematical topic, more often they do not. At the research level this will often lead to the creation of new mathematics. At the undergraduate level it leads to the reformulation of the situation in different, hopefully more tractable, mathematical terms, or to computation. Since the goal is the understanding of the original situation, resorting to computational methods or computer simulation is an acceptable alternative to developing a new mathematical theory for many applications, and for most of them at the undergraduate level. However, one ought not to turn to computation or simulation too readily. Often a deeper understanding of the original situation or the mathematical representation which has been selected leads to a fruitful approach to the mathematical problems.

The use of computation and simulation, identified as phase IV, has been discussed briefly above. Although occasionally the result of the mathematical analysis of phase III is a useable analytic expression, more frequently this result requires computation to have meaning for the original problem. For example, the result might be a theorem concerning information transmission in a complex organization. To be useful the theorem may require large amounts of data to be organized and represented cleverly in incidence matrices so as to display certain patterns. In all but trivial cases this representation cannot be carried out by hand and in order that the mathematics be useful one must develop computer programs to handle the data. It may happen that the development of a mathematical theory is quite straightforward while the invention of appropriate algorithms is very difficult and requires considerable ingenuity. The solution of these problems frequently hinges on deep understanding of the mathematics as well as skill in taking full advantage of the capabilities of the computer. The efficient use of a computer may also be an important issue. While a factor of 10 in computer time may seem unimportant from a theoretical point of view, it may make the difference between the feasibility and infeasibility of a certain approach to an industrial problem.

There is finally the task of interpreting and evaluating the results in terms of the original problem, phase V. The scientist who is applying mathematics must have sufficient familiarity with the original situation to be capable of interpreting and translating his results into the language and setting of the initial problem.

It is not uncommon to proceed through phases I-IV and find when reaching phase V that the results are not useful when viewed as statements about the original situation. In such a case one normally assumes that the mathematical analyses and computations of phases III and IV are correct and therefore looks more closely at phases I and II, especially II. We noted above that model building is crucial: an inadequate or inappropriate model may well lead through the mathematics to useless or even nonsensical conclusions. This holds independent of the quality of the mathematics used. A model based on concepts or assumptions which are deficient in some fundamental way will not produce useful results no matter how elegant the mathematical arguments used in its development.

Even if the predictions are wrong in detail, there may still have been some gain from the activity. At a minimum the situation has been examined critically with a view to identifying the underlying principles. The model building may have uncovered implicit assumptions which may or may not hold up under examination. Also, if a theory is presented in verbal form and fails, it is easy for the proponents to claim that it is essentially correct and only minor adjustments are necessary. On the other hand, if the theory is developed as a mathematical model one can more easily determine the validity of such claims. A formulation of the assumptions as individual axioms may make it easy to isolate the difficulty and modify the offending axioms instead of discarding the entire model.

3. The Construction and Use of Mathematical Models

Since the model building portion of the process described in the preceding section plays such a central role, we will consider it in somewhat greater detail. It involves the identification of certain concepts and relations as the essential ones in the situation being studied. Typically this includes an idealization and approximation of the real situation resulting in its replacement by another which is simpler in some sense but which retains the essential features of the original. Initially, the concepts singled out as fundamental may be closely identified with real things, e.g., with individuals in a population, molecules in a gas, or automobiles on a freeway, even though they may no longer be thought of as behaving strictly as real things. That is, a rat is thought of as moving instantaneously from one compartment in a maze to another, a molecule is thought of as a perfectly elastic sphere, or a population of herbivores is thought of as uniformly distributed over its range. As the model building continues the entities are viewed less as real things and more as elements in a mathematical system. The crux of modeling is the selection of an appropriate mathematical structure and useful identifications between the concepts and relations of the original situation and those of the mathematical system. Usually there is no single system which is best. There may be several systems which are natural to consider, and consequently several models for the same situation. It may happen that one of the models is distinctly preferable to the others in the sense that it accounts for the known facts and data more adequately. In such a case one normally retains that model and discards the others. There have been a number of crucial experiments specifically designed to evaluate a model or to distinguish between models. The celebrated experiment of Michelson and Morley to detect the motion of the earth through the ether and the experiments in 1960 by W. K. Estes comparing the linear model and the all-or-none model for paired-associate learning are examples. It is more common, however, that one model accounts for some observations and another model accounts for others, but that no one model accounts for all. The model(s) to be retained then depend upon which aspects of the situation are of most interest to the investigator. Among the first questions which the model builder should ask are: which aspects of the situation am I most concerned that my model include, and which have the most bearing on the detailed questions of interest to me?

Frequently a situation is studied by beginning with a simple model and cycling through the process outlined in Section 2 several times rather than by starting with a comprehensive model. The advantages of such an approach are that simple models tend to be more tractable mathematically, and that by adding assumptions one or a few at a time it may be quite clear just which assumptions lead to which conclusions. The disadvantage, and it is a very real one, is that an oversimplified model may lead to such poor predictions when compared with observations that one loses faith in the whole approach. However, a recognition of the evolutionary nature of the modeling process, and a realization that models are in general not all encompassing is usually sufficient to overcome this difficulty.

There is today no treatise which provides a definitive discussion of the theory and practice of model building. It is not even clear that such an undertaking is a reasonable one. However, there are some classifications of models, not all of them widely accepted, which provide a useful framework for comparisons. The following system, admittedly incomplete and without many desirable fine distinctions, includes several of the categories common in the literature.

(i) Models for insight and models for decisions

Model building concerned primarily with providing insight into a situation or system arising outside of mathematics has as its goal the identification of the basic processes which operate in that situation and the selection of a mathematical system and correspondences between the components of the original system and the mathematical one which illuminate the behavior of the original system. Models for decisions are designed with more specific goals in mind. Typically one is interested in making a decision or selecting a course of action so as to accomplish certain ends. The selection is to be made on the basis of the information resulting from a study of a model for the original system. Models for decision making often culminate in the development of a technique, frequently including an algorithm or simulation to be implemented on a computer, which provides solutions to specific problems or all problems of a certain class. Markov chain models for concept acquisition in mathematical psychology are usually viewed as models for insight, whereas dynamic programming models for industrial production tend to be models for decisions. In many cases an investigation has both aspects and the model is designed to provide a solid logical foundation for decision making.

(ii) Deterministic and stochastic models

Deterministic models, i.e., models based on the assumption that if there were sufficient information at one instant in time or at one stage in a process then the entire future of the system could be precisely predicted, have been widely used in the physical sciences and engineering. Stochastic models, those which describe the behavior of the system in probabilistic terms, have had their most extensive applications to situations arising in the social and life sciences. Of course, stochastic models have been used successfully in physics, e.g., statistical mechanics, and deterministic models have been used in the social and life sciences. The model of Lewis Richardson for arms races and the Lotka-Volterra models for interacting populations are examples of models arising in political science and population biology respectively which when expressed in mathematical terms lead to systems of ordinary differential equations. Many systems have been modeled in both terms. It may be that a deterministic model is selected as a first approximation to a stochastic one, and sometimes a blend of the two is appropriate. It should not be assumed that the results (or predictions) based on one type of model are necessarily better (or worse) than results based on the other. The decision as to which type of model should be constructed depends on the situation being studied and the goals of the study. It is ultimately a choice of the model builder.

(iii) Continuous and discrete models

Some situations lend themselves naturally to description in terms of continuous quantities, e.g., space or time, and others are just as naturally phrased in discrete terms, e.g., the number of automobiles produced in an hour. Even situations which initially appear to be described in terms of a continuous parameter may upon closer examination admit a natural discretization. For example, a biological population may be thought of as evolving through time. However, if observations are made periodically or if seasonal variations play a major role, then a description of the system in discrete terms may be very appropriate. There is frequently a choice and the mathematical analysis is usually quite different in the two cases. For example, difference equations may replace differential equations in a discrete model of a biological system. If computation on a digital computer is involved, then it will be necessary to return to discrete terms eventually.

(iv) Analytic and simulation models

Mathematicians tend to be better satisfied with the results of a study if their conclusions can be expressed in analytic form. This may be an analytic solution to a specific problem, or an investigation carried out largely in analytic terms and culminating in an algorithm. In either case the results can be summarized in statements recognizable as mathematical theorems. However, the complexity of actual situations may force one to admit, at least for the time being, the inadequacy of analytic techniques, and to turn to simulation. One can form a mathematical model and simulate the resulting mathematical system or one can simulate the original system more or less directly. Of course, there are various blends of analytic and simulation methods, and the simulation itself may take any of a number of forms. Psychological experiments frequently involve the simulation of an actual experience by a contrived one, interactive simulation between a human and a machine is a common training technique, and a complex engineering system such as a nuclear reactor may be simulated before it is constructed. In most instances simulation involves an analog or digital computer. There are instances in which simulation may be useful even though analytic solutions can be obtained. For example, it may be that the effect of changes in the parameters of a system on the results obtained from the model is made clearer by several simulation runs than by an examination of analytic expressions. In interpreting the results of simulation it is worthwhile to keep in mind that if a system is sufficiently complex to require simulation methods, then the model building is likely to be difficult. For instance, simulation is frequently necessitated by the inclusion of various types of random behavior in the model, e.g., random demands subject to empirically deduced probability distributions in an economic model. In such cases the results of a simulation or of several simulation runs only give estimates subject to statistical error, and the conclusions must be accompanied by some assessment of the associated random fluctuations. Although simulation is sometimes referred to as a "last resort" it is a powerful method which, if used wisely, is capable of providing information which can be obtained in no other manner.

There have been some very useful models constructed for phenomena arising in the social and life sciences. For example, probability models in genetics (Mendel's laws), logistic growth models for fruit fly populations, and input-output models in economics (for which Leontief received the 1973 Nobel Prize in Economics). Mathematical programming models have a well deserved reputation for effectiveness in business decision making. However, most of the credibility of mathematical models rests on their unusual effectiveness in the physical sciences and engineering. One could list numerous scientific phenomena whose understanding has been facilitated by the thoughtful use of mathematics. We illustrate such applications by considering what is probably the best known model of all.

4. Planetary Motion: The Evolution of a Model

The creation of a systematic explanation of the apparent motions of the planets as viewed from the earth is a major accomplishment of human intellect. The problem is one which has its origins in ancient history and also one which has attracted attention in this century. The scientific effort devoted to its study has been enormous. We will survey the subject briefly, emphasizing the modeling aspect of the various theories. There is ample literature available containing the details (e.g. [5], [11], [13], [17], [18]).

Some of the earliest attempts to explain astronomical observations adopted a view of a fixed flat earth covered by a spherical celestial dome. In the 4th century B. C. the Greeks began with such a view and devised a model which largely accounted for the data then available. This model was formulated in terms of real though somewhat idealized objects and relations. It assumed a fixed earth with a sphere containing the fixed stars rotating about it. The "seven wanderers"--the sun, moon, and five planets--moved between the earth and the celestial dome. The Greeks intended to construct combinations of uniform circular motions centered in the earth by which the movements of the seven wanderers among the fixed stars could be represented. The assumption of uniform circular motion was based on nonscientific, or at best psuedo-scientific justifications. Each body was supposed to be moved by a set of interconnecting, rotating spherical shells. Aristotle utilized the system and introduced 55 shells to account for the data available to him. In its mathematical form this model was a geometrical one. Its predictions were consistent with observations, at least to within the accuracy of the time. It was however, inadequate in two respects. First, as observations became better the model required continual refinement, and second, since each planet was assumed to remain at a fixed distance from the earth, the variations in brightness of the planets as they moved could not be explained.

Ptolemy modified this system in the second century A.D. to obtain better agreement with the data. In a simple version of his theory each planet was assumed to move in a small circular orbit (epicycle) in the period of its actual motion through the sky, while the center of this orbit moved around the earth on a large orbital circle (deferent). The deferent and epicycle alone were insufficient to account for the observed irregularities in planetary motion, and Ptolemy also introduced the equant, an axis of uniform motion off center within the deferent. Only from this position would the planet appear to move with uniform angular velocity. The earth was assumed to be off center in the opposite direction from the equant with respect to the center of the deferent. This system provided adequate flexibility for the planets other than Mercury; with a slight modification it would serve for Mercury also.

Early in the 16th century Copernicus became dissatisified with the Ptolemaic equant, which seemed to him to violate the principle of uniform circular motion, and he proposed modifications involving more epicycles, off center deferents, and finally a heliocentric (sun centered) system. It is interesting that his early defense of the heliocentric system is based entirely on esthetics, particularly on a plea for simplicity. He proposed that the earth and the other planets revolved around the sun in uniform circular orbits. Since his model retained the assumption that the basic motions of the planets were circular, it was also necessary to retain the epicycle concept to account for variations in the brightness and apparent velocity of the planets as viewed from the earth. This model was based on geometry, as was Ptolemy's, and not on physics. Indeed, the editor of one of Copernicus' major works writes "But these hypotheses need not be true or even probable," and "If they provide a calculus consistent with the observations, that alone is sufficient."

Toward the end of the 16th century a Swedish astronomer, Tycho Brahe, collected masses of detailed observations on the motions of the planets. Tycho seems to have viewed Copernicus as a builder of hypothetical geometric models, and, in an attempt to formulate models incorporating the apparent physical reality of a sluggish, massive earth, he proposed another geocentric model. It is not his model for which he is accorded his status in astronomy, however, it is for his systematic observations which provided the foundations for future work. Johannes Kepler inherited Tycho's records and he undertook to modify Copernican theory to fit these observations. He was particularly bothered by the orbit of Mars whose large eccentricity made it difficult to fit into a deferent - epicycle system. He was eventually led to make a very creative step, a complete break with the Platonic-Pythagorean uniform circular motion hypotheses that had so dominated astronomy. He posed as a model for the motions of the planets the following three assumptions, usually referred to as Kepler's laws:

1. The planets revolve around the sun in elliptical orbits with the sun at one focus (1609).

2. The radius vector from the sun to a planet sweeps out equal areas in equal times (1609).

3. The squares of the periods of revolution of any two planets are in the same ratio as the cubes of their mean distances from the sun (1619).

These empirical laws are simply statements of observed facts. They are, however, perceptive and especially useful formulations of the regularities noted in the observations. Along with the identification of these laws, Kepler hypothesized a physical mechanism, a force emanating from the sun, which accounted for the motion of the planets. This model accounted very well for the accumulated observations and set the stage for the next refinement.

The models developed up to the middle of the 17th century had an empirical or geometrical basis with a minimum of support from physics. Isaac Newton's theory of gravitation provided simultaneously a physical interpretation and a concise and elegant mathematical description of cosmological phenomena. The combination of the laws of motion and the universal law of gravitation furnished a mathematical system from which the motions of the planets could be deduced. In this setting the motion of a planet could be determined by first considering the two-body system consisting of the planet and the sun. The motion of this system was easy to determine, and the results were the three laws of Kepler. These predictions are good first approximations since the sun is the dominant mass in the solar system and the planets are widely separated. However, according to the law of gravitation, each planet is also subject to forces due to each of the other planets, and these forces result in perturbations of the elliptical orbits predicted on the basis of the two-body model. A careful examination of the orbit of a specific planet resulted in the identification of perturbations due to each of the remaining known planets. If these pertubations did not account for the total of the observed deviations from the behavior predicted on the basis of a two-body model, then one might try to account for the remainder by assuming the existence of a yet unknown planet. Estimates on the size and location of the hypothesized planet could be obtained and a search initiated. In fact, this is the sequence of events which resulted in the discovery of Uranus, Neptune and Pluto. It is an impressive triumph for the mathematical system proposed by Newton that minute discrepencies between theory and observations could lead to the discovery of unknown and in fact unanticipated planets.

Even the remarkable model of Newton does not account for all the data and there has been further refinement in this century. Small perturbations in the orbit of Mercury, unexplainable in Newtonian mechanics, provided some motivation and support for the development of the theory of relativity. The modified version of the Newtonian theory incorporating relativistic corrections appears to be adequate for existing data. New data or revised interpretations of existing data may of course necessitate further revision.

Any of the models mentioned here can be formalized. That is, one can identify the basic concepts which are important, e.g., force, mass and position, select some as undefined terms and provide precise definitions of the others. The resulting mathematical system can be studied as an abstract structure without relevance to its origin or to any possible real world meaning of the terms.

The attempts to account for the motion of the planets described here illustrate several aspects of model building. First, and perhaps most vividly, the typical cycling through the model building process is demonstrated. Some of the successive refinements are clearly defined major departures from previous efforts, while others are simply minor modifications. These revisions were initiated for varying reasons. Copernicus was primarily concerned with simplifying the Ptolemaic model while Newton was interested in finding physical and mathematical principles from which Kepler's laws (among others) could be deduced. The search for a simple model based on a few easily understood principles is characteristic of modeling. Also, in addition to modifications based on a desire for simplicity or elegance, the refinement of a model necessitated by new data is illustrated. What has not been discussed, other than the brief mention of the geocentric model of Tycho Brahe, are the many dead ends and useless models which were considered in the process of discovering useful models. Some such efforts have been recorded in the literature, many more have been lost.

5. An Example from Psychology

The model for physical phenomena known as classical or Newtonian mechanics, which provided one of the models sketched in the preceding section, has several desirable characteristics. First, the mathematical laws are simple and their relation to well understood physical concepts is clear. Next, the model is quite comprehensive and it can be used to study an enormous variety of situations in physics and engineering. It is unnecessary to construct a number of variations of the model to account for the results of different kinds of experiments. Thus, to cite a simple example, the motion of a cylinder rolling down an inclined plane can be described using the same basic principles as are used to describe the motion of a water droplet leaving a garden hose. Finally, there is (nearly universal) confidence in the results or theorems which follow from an analysis of the model. It is accepted that the model does indeed illuminate the underlying physics of the situation.

On the other hand, the models constructed for use in the social and life sciences and in business do not in general have this simplicity and comprehensiveness and they have sometimes inspired less confidence in their users. Typically the basic concepts and mathematical relations may be relevant only to a rather restricted set of situations, and the degree to which they illuminate the underlying scientific or business problem may be less than totally clear. In addition, problems of parameter estimation may severely restrict the utility of otherwise promising models. It is, of course, by no means the case that models in the physical sciences are always simpler and more useful than models in the social and life sciences. For example, relatively simple mathematical models have proved effective in genetics ([7], [8]) and in learning theory. Applications of mathematics to learning theory are especially interesting because of the variety of models which have been developed to account for the experimental evidence. It is useful to comment briefly on this field since it illustrates the development of models very different from those of mechanics.

Paired-associate learning is one of the topics included under the general designation of verbal learning. In a typical paired-associate learning experiment a subject is presented with a list of stimulus-response pairs, one pair at a time. For example, one pair in the list might be XW-4. Here XW is the stimulus member and 4 is the response member. Depending upon the particular experimental design being used, the subject might be presented with a stimulus and given a short time to respond before the associated response member appears. This routine is repeated for each item on the list. A trial is one presentation of the entire list. Usually the order of the items is determined randomly for each trial. The experiment will proceed either for a fixed number of trials or until a criterion level is reached. For example, the criterion level might be two successive errorless trials.

A major objective of a learning theory is to confirm or predict a learning curve, a measure of performance as a function of time. For paired-associate learning one usually takes the measure of performance to be the proportion of correct responses and the time as the trial number. This objective is usually not sufficiently refined and several reasonable models with two or three free parameters can be made to agree adequately with empirical learning curves by a careful selection of parameter values. Consequently, one normally has to use other methods of selecting between alternative models. For example, one might base a distinction between models on a detailed analysis of response sequences.

We mention the basic ideas of two models to indicate the possibilities. A detailed discussion of these two models is included in [1], [3], and [15]. Alternative models are proposed in the references just cited and in [16] and [20]. Bush and Mosteller compare eight of the common learning models in [2].

Of the many versions of the linear model, the following is probably the most elementary. Consider a specific item in the list and let P_n be the probability that the subject makes a correct response to that item on trial n. It is reasonable to suppose that P_{n+1} depends in some way on P_n. Various rationales (see [3] p. 51 and [15] p.3) can be given to support the assumption that

$$P_{n+1} = \alpha A + (1-\alpha) P_n,$$

where A is the asymptotic value of P_n for large n, and α is a parameter which must be estimated from the data.

One of the alternatives to the linear model is the all-or-none (or one-element) model. The view of the learning process which underlies the all-or-none theory is very different from that of the linear model. In a simple version of the all-or-none view of learning the subject is supposed to be in one of two states, the learned state (L) or the unlearned state (U). During each trial the subject either remains in L or U or makes a transition from U to L. It is assumed that once the subject reaches state L it remains there. If the subject is in state L, it is supposed to make only correct responses while if it is in state U it may or may not make correct responses by chance. Finally, it is assumed that the subject begins in the unlearned state U. In one direction these assumptions can be given an intuitive rationale and in the other direction they can be made quite precise (see Section 3.1 of [1]).

It is natural to take a Markov chain as a mathematical system whose structure is appropriate for the all-or-none model. A Markov chain model of the simple version described above has a transition matrix of the form

$$\begin{array}{c} \\ L \\ U \end{array} \begin{array}{c} \begin{array}{cc} L & U \end{array} \\ \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} \end{array}$$

where c is a learning parameter. It is not learning but responses which are observed in experiments, and the connection between state occupancy and response is given by

$$\Pr[\text{correct response} \mid \text{state L}] = 1,$$

$$\Pr[\text{correct response} \mid \text{state U}] = \beta,$$

where β is a response parameter. It is customary to estimate β in advance from the number of response alternatives available to the subject. Thus the predictions based on this model will be functions of the single parameter c which must then be estimated from the data.

The predictions of these two rather different models are analyzed and compared in [1], Chapter 3, and [15], Chapter 2. In the linear model learning is viewed as a change in response probability from one trial to the next. Thus each presentation of an item is assumed to increase the probability of a correct response. On the other hand, as its name implies, the all-or-none model is based on the assumption that the effect of a single presentation is either to produce complete learning of the association or no learning at all. For a specific set of data, it is reasonable to expect that one should be able to choose between the models based on these rather different hypotheses. This is the case in the comparisons cited above.

6. Concluding Remarks

There are two points to which we return for emphasis. The first relates to the criterion of simplicity frequently mentioned earlier in this paper. The idea is an imprecise one, and the degree of simplicity or complexity ascribed to a model is frequently dependent upon the observer. However it usually happens that when a situation is modeled in full generality the resulting mathematics is sufficiently peculiar to the situation that one encounters difficulties in taking advantage of known results. The alternative then to developing a new mathematical theory is to make simplifications which bring the mathematical systems arising from the model within the scope of known theories. These theories may have been developed in the analysis of a model created in the study of another situation. This brings us to the second point. One of the major advantages of using mathematics to describe scientific phenomena is that its use enhances markedly the possibility of recognizing similarities between situations which may appear superficially quite different. These similarities may allow the use of results which were derived in the analysis of a model constructed for one purpose to study a very different situation. For example, results on linear inequalities derived by J. Frakas in his study of an engineering problem have proved to be very useful in studying linear programming problems which arise in economics and business. The recognition of a common mathematical structure may even lead to the identification of new scientific principles. This is exemplified in the biological least-action principles identified by noting the similarities in models for biological systems and those for physical systems ([6], [9], [17]).

We conclude with a comment which is somewhat out of place in this article but entirely appropriate for the volume. The repeated successes of certain individuals in creating and developing mathematical models for scientific phenomena indicates quite clearly that model building can be learned. It is not as clear that it can be taught in the same way as most academic subjects. The novice needs to be an observer initially and then, as quickly as possible, a doer. A beginner benefits considerably by observing the process of applied mathematics in action. To this end, it may be that examples which are not original but which are new and meaningful to the student serve as well as new examples created for the purpose. However, modeling is not learned by watching others build models, even if this includes some of the false starts and failures that are so common, but rather by becoming actively and personally involved in the modeling process. To be sure, there will be frustrations and futile efforts. But only through participation can one gain understanding and facility in applying mathematics.

References

The references provided in this list are cited in the paper. Those selected for examples in sections 4 and 5 were chosen because they give attention to the model building aspect of the work.

[1] Atkinson, R. C. G. H. Bower, and E. J. Crothers. An Introduction to Mathematical Psychology. New York, John Wiley and Sons, 1966.

[2] Bush, R. R. and F. Mosteller. "A Comparison of Eight Models," in Studies in Mathematical Learning Theory, eds. R. R. Bush and W. K. Estes. Stanford, California, Stanford University Press, 1961.

[3] Bush, R. R. and F. Mosteller. Stochastic Models for Learning. New York, John Wiley and Sons, 1955.

[4] Coulson, C. A. The Spirit of Applied Mathematics. Oxford, Clarendon Press, 1953.

[5] Gingerich, O. "Copernicus and Tycho," Scientific American, 229 (1973).

[6] Goel, N. S., S. C. Maitra, and E. W. Montroll. Nonlinear Models of Interacting Populations. New York, Academic Press, 1971.

[7] Karlin, S. and M. Feldman. "Mathematical Genetics: A Hybrid Seed for Educators to Sow," Int. J. Math. Educ. Sci. Technol., 3 (1972), 169-189.

[8] Karlin, S. "Some Mathematical Models of Population Genetics," Am. Math. Monthly, 79 (1972), 699-739.

[9] Kerner, E. H. Gibbs Ensemble: Biological Ensemble. New York, Gordon and Breach, 1972.

[10] Klamkin, M. S. "On the Ideal Role of an Industrial Mathematician and its Educational Implications," Am. Math. Monthly, 78 (1971), 53-76. See also the many references provided with this article.

[11] Koestler, A. The Sleepwalkers. New York, Macmillan Company, 1959.

[12] Lin, C. C. "Objectives of Applied Mathematics Education," SIAM Rev., 9 (1967), 293-311.

[13] Pollak, H. O. "Applications of Mathematics," in Mathematics Education. Chicago, Illinois, University of Chicago Press, 1970.

[14] Ravetz, J. "The Origins of the Copernican Revolution," Scientific American, 215 (1966).

[15] Restle, F. and J. G. Greeno. Introduction to Mathematical Psychology. Reading, Massachusetts, Addison-Wesley Publishing Company, 1970.

[16] Restle, F. *Mathematical Models in Psychology*. Middlesex, England, Penguin Books, 1971.

[17] Samuelson, P. A. "A Biological Lest-Action Principle for the Ecological Model of Volterra-Lotka," *Proc. Nat. Acad. Sci. USA*. 71 (1974), 3041-3044.

[18] *The Scientific World of Copernicus*. ed. B. Bieńkowska, forward by Z. Kopal. Dordrecht, Holland, D. Reidel Publishing Company, 1973.

[19] Small, R. *An Account of the Astronomical Discoveries of Kepler* with a forward by W. D. Stahlman. Madison, Wisconsin, University of Wisconsin Press, 1963.

[20] Sternberg, S. "Stochastic Learning Theory," in *Handbook of Mathematical Psychology*, Vol. II, eds. R. D. Luce, R. R. Bush, and E. Galanter. New York, John Wiley and Sons, 1963.

[21] Wigner, Eugene P. "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," *Comm. Pure and Appl. Math*. XIII (1960), 1-14.

The references listed below, while not cited in the paper, provide additional sources of information and examples. Some of the items are concerned with applications to particular situations or classes of situations, and others illustrate and study the model building process.

Apostel, Leo. "Towards the formal study of models in the non-formal sciences," in *The Concept and The Role of the Model in Mathematics and Natural and Social Sciences* (H. Freudenthal, ed). New York, Gordon and Breach, 1961.

Bailey, Norman T. J. *The Mathematical Approach to Biology and Medicine*. New York, John Wiley and Sons, 1967. See especially Chapter 3, "The Process of Scientific Research."

Ball, R. J. "Econometric Model Building," in *Mathematical Model Building in Economics and Industry* (M. G. Kendall, ed). New York, Hafner Publishing Company, 1968.

Bergstrom, A. R. *Selected Economic Models and their Analysis*. New York, American Elsevier Publishing Company, 1967. See especially Chapter 1, "Introduction."

Cohen, Hirsh. "Mathematics and the Biomedical Sciences" in *The Mathematical Sciences*. Cambridge, Massachusetts, The MIT Press, 1969.

Dantzig, George B. *Linear Programming and Extensions*. Princeton, New Jersey, Princeton University Press, 1963.

Dyson, Freeman J. "Mathematics in the Physical Sciences" in *The* Mathematical Sciences. Cambridge, Massachusetts, The MIT Press, 1969.

Kemeny, J. G. and J. L. Snell. Mathematical Models in the Social Sciences. Waltham, Massachusetts, Ginn/Blaisdell, 1962. See especially Chapter 1.

Kendall, M. G. "Model Building and its Problems," in Mathematical Model Building in Economics and Industry (M. G. Kendall, ed). New York, Hafner Publishing Company, 1968.

Kowal, Norman E. "A Rationale for Modeling Dynamic Ecological Systems," in Systems Analysis and Simulation in Ecology (Bernard C. Patten, ed). New York, Academic Press, 1971.

Luce, R. D. and H. Raiffa. Games and Decisions. New York, John Wiley and Sons, 1957.

Maki, Daniel P. and M. Thompson. Mathematical Models and Applications. Englewood Cliffs, New Jersey, Prentice-Hall, 1973. See especially Chapter 1, "Basic Principles."

Rapoport, Anatol. "Lewis F. Richardson's Mathematical Theory of War," Journal of Conflict Resolution (1) 1957, 244-299.

Saaty, Thomas L. Mathematical Methods of Operations Research. New York, McGraw-Hill Book Company, 1959. See especially Chapters 1 and 3.

Saaty, Thomas L. Topics in Behavioral Mathematics. Washington, D. C., Mathematical Association of America, 1973.

Simon, Herbert A. "Some Strategic Considerations in the Construction of Social Science Models," in Mathematical Thinking in the Social Sciences (Paul Lazarsfeld, ed). Glencoe, Illinois, The Free Press, 1959.

Suppes, Patrick. "A comparison of the meaning and uses of models in mathematics and empirical sciences," in The Concept and The Role of the Model in Mathematics and Natural and Social Sciences (H. Freudenthal, ed). New York, Gordon and Breach, 1961.

Several of the NSF Chautauqua-Type Short Courses (sponsored cooperatively by AAAS and NSF) have Study Guides which contain examples of mathematical modeling. Those courses offered during 1974-75 for which study guides were prepared are listed together with the lecturer(s) below.

Behavior-Genetic Analysis, Jerry Hirsch.

Patterns of Problem Solving, Moshe F. Rubinstein.

Public Policy Analysis: Theory and Some Applications, Elinor Ostrom.

Water Polution, David Kidd.

Conflict Regulation, Paul E. Wehr.

Mathematical Modeling and Computing in the Physical, Biological, and Social Sciences, William Dorn and Jack Cohen.

Atmospheric Sciences, Vincent J. Schaefer and Volker A. Mohnen.

Chapter 2
THE TRIAL TEACHING OF THE MODULES

Introduction.

Individuals and committees in the mathematics community have been insisting for a long time that students of mathematics need new kinds of courses added to their programs to provide realistic experiences in the applications of mathematics. While other proposed reforms in the mathematics curriculum have gone forward just in response to such urgings, in particular to various sets of recommendations made by panels of CUPM, the creation of the desired kind of courses in applications has proved to be a most intractable problem.

It was, consequently, a primary objective of the present project, not only to assemble sample materials to facilitate such a course, but to carry through its implementation on a trial basis and report the results, for the guidance that might thus be provided to others.

Seven teachers were identified at an early stage in the project, chosen from institutions thought to be typical of the kind that would implement such a course. The main special qualification they had in common was their willingness to try teaching a course of this kind. Their names and institutions are as follows:

1. Thomas S. Angell, University of Delaware
2. Charles A. Hall, University of Pittsburgh
3. Benjamin D. Haytock, Allegheny College
4. Henry E. Heatherly, University of Southwestern Louisiana
5. Roger H. Pitasky, Marietta College
6. Kenneth R. Rebman, California State University, Hayward
7. Kenneth Stofflet, University of Evansville

Preliminary drafts of the modules were supplied to each teacher and he was requested to indicate which ones he preferred to use and whether he intended to teach it in a traditional fashion or in an open-ended way, one that would involve the students in the development of a model beginning with only general background information on the problem. These preferences were, perhaps surprisingly, diverse, with the result that only a few second preferences had to be used.

When the classes were completed, a conference was held, at which the teachers presented reports of their experiences, the authors heard suggestions for improving their materials and an informal dialogue took place. The versions of the modules included in this report incorporate many changes resulting from these communications. (The modules "A Model for Municipal Street Sweeping Operations" and "A Mathematical Model for Renewable Resource Conservation" were incorporated into the project after the trial teaching stage began.)

The summaries that follow attempt to impart the highlights of each teacher's experience, by abstracting and paraphrasing his written and oral reports. They include basic factual information: the nature of the institution, the kind of students in the class, the material covered and the method of coverage, any use of the computer, tests, projects and supplemental readings, etc. In addition, those comments that best reveal the reactions of teachers and students to the experimental course are stressed. A particular effort is made to throw light wherever possible on the question whether such material is best taught in a traditional manner, or through some form of open-ended instruction.

The reports submitted by the teachers and the notes of their oral presentation (and the discussions) at the conference are the sources from which the following summaries have been derived. It should be mentioned that this material has been edited and paraphrased liberally in an attempt to convey the individuals' personal experiences faithfully, if briefly. Thus, although mostly cast in the first person, much of what follows is not directly quoted from the individual in question.

Finally, a number of observations made in the individual reports found substantial endorsement by the group assembled at the conference. These have been considered worthy of separate listing at the end by way of general guidelines to those who contemplate teaching the kind of course in question.

Summaries.

1. Thomas S. Angell, University of Delaware

My reaction, and the reaction of the students was a very mixed one. Both they and I have been more accustomed to the well-defined progression of topics and techniques which characterize the staples of the current curriculum. We all found this seminar course much different from our usual fare and the end result was rough rather than polished.

Some very good things did happen, primarily on an individual level, which made the seminar worthwhile for the students. For example, one student, working on a term project analyzing the voting power in the Faculty Senate, with respect to new proposals changing the current representation, has had first-hand experience with the usefulness of a well-designed algorithm for computation. The dramatic savings in computer time, gained after we discussed his computer problems and studied the procedures outlined in the module, will be remembered for a long time. The course was even inspirational, if one takes student comments on the post-course questionnaire at face value: Because of [this course] I'm trying to get into graduate school to major in Operations Research.

We did not have an unqualified success at Delaware. My intention was to conduct a seminar in which students would share the responsibility for presentation of the material, including reports on outside reading as well as the material and exercises contained in the modules. The idea was to distribute parts of the modules to teams of students, to meet with them outside of class for guidance and preparation, and then to have them present the bulk of the material to the class. Indeed, we began with this format, but, as the semester progressed, it was necessary to modify the procedure. The most commonly voiced criticism at midterm was the feeling that we were going too slowly. My assessment is that the students would have preferred a faster pace, but that was not possible, at least in part because of the degree to which students were presenting the material--not a surprising result.

The department at Delaware is very committed to applied mathematics and, in particular, differential equations. This course attracted some of the better students. The course was taught under the title "Undergraduate Topics in Applied Mathematics," which has run for several years but has been primarily of the classical variety dealing with topics in fluid dynamics, methods of mathematical physics and tensor analysis. This is the first time anything along the lines of modeling has been taught in the mathematics department. There is a course along these lines taught in the engineering department and several of the students had either audited or taken that course.

The initial registration was eight seniors, six juniors, and eight sophomores. Disciplines represented were Mathematics (10), Statistics and Computer Science (8), Electrical Engineering (2), Biology (1), and Physics (1). In the course of the semester, we lost one senior and two sophomores, so the ultimate class size was nineteen. With this distribution of students, one of the recurring problems was the wide divergence of background which led, ultimately, to more lecturing than originally planned.

The seminar met once each week on Monday nights for a three-hour session. This is clearly not an optimal schedule, but necessary for this group due to other scheduling problems. While there are obvious disadvantages in this format, there are advantages too, not the least of which is that it is very different from the usual class format and, as such, created a strong feeling of group cohesion.

The course began with background material on modeling, then took up the general question of population growth in connection with the Ecosystems module. Eight weeks were spent on this, using an open-ended approach: reports on references; a study of a single population using difference equations; basic ideas of two dimensional systems of differential equations; work on the module itself.

In addition to the module, we had reports from students on a paper by Huffaker which discusses the problem of exhibiting sustained oscillations for a prey-predator system in the laboratory and some experiements of Gause. A problem with presenting material on this subject was the lack of field data. Quantitative information of the type suitable for comparison with a computer run seems hard to come by. A major source of student dissatisfaction with this and other modules was our inability, due either to lack of scientific background or hard data, to criticize the models we worked with. The assumptions that the models were based on were, I think, clearly set out during the course. How well the model reflects observations was not as clear, and student reaction to this module was that the models seemed artificial and the problems primarily directed to mathematical problems rather than to modeling issues.

During this time, one class session and part of another was devoted to a team brainstorming approach to a simple epidemic model. I added this activity because the students felt that they were studying the process of model building as practiced by other, and not doing their own modeling. When they did attempt to formulate models they tended to come up with exceedingly difficult ones. One of the problems was the lack of time to follow through and refine some of these models.

These sessions were popular with the students and did much to get them involved in the course. Difference equations were chosen by all but one of the six teams for their models, mostly including non-linearities. Time pressures prevented following up these sessions with an analysis of the models, which limited the value of these sessions.

The next session (three weeks) focused on the module on weighted voting systems. Both conversations with students and written comments indicate that this was the most popular of the materials studied. Its popularity (nine students ultimately chose projects on the subject) seemed to derive from the character of the mathematics involved. To most of the students, this was new, in contrast to the other two, which dealt with differential equations.

We began work on this subject with a lengthy class discussion to focus their perception of the problem. This was followed by my presentation of the political science background. Breaking into small groups, the students were quite successful in formulating a method of measuring power--essentially the Shapley-Shubik concept. It was then possible for me to formalize the concept following Lucas' presentation.

The last sessions were devoted to the module on permafrost. Few of the students had any acquaintance with partial differential equations and none had seen the Stefan problem. I chose to adopt a strictly lecture presentation. My own mathematical interests make this module particularly appealing and my own opinion is that it is well constructed. The students found this the most difficult mathematical material, although the physical problem as described in the module is very clear.

In conclusion, I think that the students, for the most part, found the course stimulating and rewarding, although it did not live up to all of their expectations. As the instructor, I felt it important to give the students a lot of background information and fulfilling this obligation required many hours of preparation. I found myself spending hours in the library, many more than with any other course I have taught. It was enjoyable, but I think that instructors in the future might find it helpful to have well selected background material identified for their use, and that they be urged to start their studies early. If the course is to be taught in an open-ended manner, two modules could be easily covered, but three would be a little ambitious.

The problems discussed are interesting to both student and instructor. They all have an aura of relevance and importance although, as mentioned earlier, the lack of data is limiting. In all of the problems, I tried to emphasize that we were dealing with serious problems and not cookbook examples. The students responded with a mixture of enthusiasm and anxiety that it might not be possible to get good results. If I were to teach this course again, I would put more time into computer testing of the models and it would be helpful if necessary data could be supplied. Many of my students had the capability to use the computer and several did in their term projects. More use of the computer would have been very useful in rounding out the presentation.

I think that the students' comments showed that the general idea of the course, and the subjects presented, were enthusiastically received, and well worth the time and energy spent.

Selected Student Report Topics - University of Delaware

1. A markov Chain Model for Prisoner's Dilemma.
2. A Free Boundary Problem of Gas Diffusion.
3. Traffic Flows for Newark, Delaware; Graph Theory Models.
4. Application of the Static Helmholtz Equation to a Spherical Reactor.
5. Dynamic Programming Approaches to Optimal Vaccination Schemes.
6. Permafrost Model Including Convection.
7. Analysis of the Canadian Constitutional Amendment Scheme.
8. Electoral College Reform.
9. Computer Redistricting - Analysis of the CROND Scheme.
10. Analysis of Voting Power in the Faculty Senate, University of Delaware.
11. A Mathematical Model of Schistosome Infections in Man.

2. Charles A. Hall, University of Pittsburgh

The course was taught using an open-ended technique. Students were exposed to the five stages in the "evolution and dispatch" of an industrial problem (as described by Pollak and Klamkin): Recognition, Formulation, Solution, Computation and Explanation, in connection with the module on Steam Generators.

This course, "Industrial Mathematics," has been taught the past four terms and is designed "to simulate an industrial environment."*

There were a total of seven students completing the course (two additional students withdrew during the second week of the course): six seniors and one junior; all seven were mathematics majors; they averaged 33 credit hours in mathematics and 12 credit hours in computer science.

---

* See C. A. Hall, Industrial Mathematics: A Course in Realism, American Mathematical Monthly, vol. 82, no. 6, June-July, 1975, pp. 651-659.

Two students used the experience of the course to good advantage in interviews with prospective employers.

Student reaction was very enthusiastic for the most part. They did seem to sense the inherent connection between such an analysis and the "energy problem." I strongly recommend the "open-ended" approach for a change-of-pace course at the senior level. The students were most interested in applying (even if in a somewhat limited sense) what they had learned in their other mathematics and computer science courses; the unstructured open-ended approach allowed them a new flexibility to try out their expertise (or lack of it) in problem solving. Students were not as timid as one might expect to ask questions of me and each other as the term proceeded.

The class worked as a team to write one computer program STEAM with each member being given explicit tasks as the term proceeded. Some students contributed more to the team effort, did better on explicit tasks and demonstrated better overall understanding of what we were doing: they were graded accordingly.

It was made clear on the first day of class that their grades would be determined by my evaluation of their performance as problem solvers.

Each student was asked to write an individual progress report in the tenth week which was helpful in evaluating his/her contributions to date.

The prerequisites were calculus and linear algebra. However, the presentation of the material is greatly enhanced if one is contemplating implementation on a computer; some of the more interesting discussions were initiated by the attempt to implement the algorithms.

I stressed the extensive use of the digital computer in the modern day industrial world in solving problems of Physics and Engineering. Our goal was the implementation of three numerical algorithms for solving the nonlinear network equations. Even though the program STEAM is not error-free, it was a vehicle by which the class endeavored to understand the problem of modeling fluid flow in a steam generator. They seemed to take pride in their limited success and definitely benefited from implementing their ideas.

The following is a broad description of how our class spent the 14 weeks on this module and related material. The course met twice a week, Tuesday and Thursday (3:00-4:20), and students earned three academic credits for completing the course.

Class # 1. Discussion of industrial mathematics

2. No class

3. Process of applying mathematics and problem solving

4. Introduction to Network Analysis Module - Origin of the problem

5. Hand-out (Nuclear Power and the Environment) - Network Theory

6. Network Theory

7. Network equations - derivation

8. Uniqueness theorem-proof
9. Modeling the flow problem
10. Jacobi and Gauss-Seidel methods for linear models
11. AEC films
12. Outlined the input of STEAM - Formation of network equations
13. Numerical solution of network equations
14. Nonlinear SOR - program strategy
15. Work period
16. Review program strategy - progress reports - oral
17. Bisection - regúla falsa - Newton's method
18. Program debugging - progress report - written
19. Technical writing - program debugging
20. Assignments on report - program debugging
21. Program debugging - Bisection method
22. Work period
23. Video-taped discussions with Professor Porsching
24. Analysis of computations
25. Report writing - analysis of computations
26. Review of problem and its solution

The class wrote and typed a "Technical Report" which described their interpretation of the problem, their solution techniques and program strategy. Several students commented that having to "put-it-all-together" in writing was valuable to them since they had to go back over their notes and towards the end of the term they began to appreciate better what was going on.

3. Benjamin D. Haytock, Allegheny College

Allegheny College is a small (1800 students) liberal arts college. We do not offer any advanced degrees nor do we have an engineering program.

These modules were used in a selected topics course at the junior-senior level. There were seven students in the class--all mathematics majors. Only one had a good background in physics but all were good to very good students. All but one were fairly proficient in FORTRAN. All had had ordinary differential equations and all but one had had linear algebra.

The student reaction to the course was very positive. They all said they worked hard--harder than in most courses--but that the course was worth the extra effort. All felt that the course should be made a part of the regular curriculum.

I taught the modules on Heat Transfer and Helminth Infections. My colleague, Dick McDermot, taught the Steam Generator module.

The three modules were presented in a traditional lecture manner. If I was doing the material over again, I think I would use an open-ended approach to the Helminth Infections module. I do not think an open-ended approach to the Heat Transfer or Steam Generator modules would have been fruitful in my class.

Students were evaluated on the basis of assigned problems, two computer programs, and an oral final examination.

Heat Transfer in Frozen Soil was the first module we took up. This is because I wanted to have as much time as possible to spend on it. The problems in the module were assigned and graded, and a program to implement the algorithm was also assigned. For this last, the students were divided into three groups and each group wrote their own program to handle both a two and a four layer problem. This worked very well, and the students seemed to benefit from working in small groups. It was a time consuming project, but it helped the students in understanding the details of the algorithm. All felt that it was worth doing.

The students found this problem very interesting--difficult to understand at first--but all had a good grasp of the main ideas by the time we had finished. I plan to use it again.

The Steam Generator module was the second we covered (it was taught by Dick McDermot). As with the first, a program was assigned to implement the method. The problem was to duplicate the results of the sample problem contained in the module. The students worked in groups again and were all fairly successful. The only problem we encountered was with the speed of the IBM 1620 and not being able to follow full convergence. Since the flow calculations are quite sensitive to errors we had some discrepancies there.

Problems were not provided, but we made some up and assigned them for credit also. These included some small network problems that could be set up and done by hand and a convergence proof for a two by two Gauss-Seidel iteration.

The discussion of the basic assumptions and factors not taken into account in developing the link characteristic was useful. Questions had been raised in class about these things, and the students were glad to see some of their objections confirmed.

The students liked the module but were unable to see how to use the computational results (pressures and flows) to answer the original basic questions dealing with corrosion, the life expectancy of the generator, and how to use this information to design a good generator. Material dealing with these aspects of the problem should be included. We were unable to help the students on this.

The most striking aspect of this problem for the students was the application of network theory to this type of problem. This was unexpected and seemed to impress them.

In some ways, MacDonald's Work on Helminth Infections was my favorite of the three modules covered. The significance of the problem is easily grasped and the students were all impressed by the way mathematics could be used to help in solving a problem of this nature.

A second nice feature is that most of the module could be developed through an open-ended approach if desired.

All in all I thought the course was very successful and both the specific modules and especially the idea behind them very good.

[Note: After the conference, and as a result of his work on the Helminth Infections module during the course, Professor Haytock was asked to help in the writing of the final version of the module and is now listed as a co-author.]

4. Henry E. Heatherly, University of Southwestern Louisiana

The course was offered as a senior level, three credit hour, mathematics course, run for the most part on a seminar basis. Prerequisites were a three semester calculus sequence, a semester course in elementary differential equations, and permission of the instructor. Five students took the course: three senior mathematics majors, one graduate (M.S. level) student in applied mathematics, and a junior in computer science. Each had a strong mathematics background (considerably exceeding the prerequisites) and high ability. Every student could program in FORTRAN and knew how to use the local computer facilities; two were expert programmers. The mathematics majors had done well in courses in abstract algebra (Herstein), advanced calculus (Gaughan), and linear algebra (Hoffman and Kunze).

The students were highly enthusiastic about the course and enjoyed it immensely. They felt it to be an important course that should be made part of the curriculum in mathematics. Two of the mathematics majors changed their plans for graduate work after taking the course and have decided to do work in applied mathematics. The students stated that the modules we studied seemed important and relevant.

We took up the modules in the following order: (1) On Population Mathematics, (2) MacDonald's Work on Helminth Infections, (3) Measuring Power in Weighted Voting Systems, (4) Heat Transfer in Frozen Soil. The latter was not treated fully because of the amount of mathematical background it required.

The computer played an important role throughout the course. The students gained new insight into the difference between an analytic solution and actually obtaining numbers one could apply to the real world. They also had their eyes opened to the difficulties involved in obtaining useful data and interpreting data and solutions.

Conversation with colleagues was helpful at times. Two instances in particular should be noted. A colleague in statistics was very helpful in explaining and giving insight into why certain random variables in the Helminth Infections module have (or might reasonably be assumed to have) a Poisson distribution. A colleague in physics was helpful in discussing heat transfer.

The students were evaluated on the basis of:

1. problems or projects worked outside of class (individually or in groups) and presented in class (orally),

2. oral reports to the class on outside reading and study,

3. class participation in the discussions,

4. written reports done individually outside of class,

5. problems worked outside of class on an individual basis and written up to be turned in.

There were no quizzes or exams. This now seems to have been an error and the next time I teach a course of this type I will give tests, perhaps a quiz over each module. For a larger, more heterogeneous class quizzes would seem to be imperative.

The initial module for the course as selected by the students was On Population Mathematics. Their interest was high and they launched into reading the module with gusto. They found the module somewhat unclear and far too sketchy. The latter led us to hunt for other sources of information as well as to attempt to build our own models. I found I had to supplement the material considerably and to explain parts of the module and the reference material on the renewal equation.

The students and I agree that the module on MacDonald's Work on Helminth Infections is well written, easy to read, and a stimulating example of mathematics successfully used in a real world situation. The students wanted to know more about the work being done on the subject so we wrote to Professor Hirsch at the Courant Institute. He sent us several papers and the lengthy report: "Mathematical Models of Some Parasitic Diseases Involving an Intermediate Host," Ingemar Nåsell and Warren M. Hirsch. Some of the students read parts of this report. We also found it very useful to read pertinent sections on parasitic diseases in several books on tropical diseases. One of the students made a report to the class on biological-medical features of the problem after doing outside reading; this helped in appreciating and criticizing the models studied.

When we took up Measuring Power in Weighted Voting Systems the change of pace from biological problems to social-political ones and from analytical methods to combinatorial methods was refreshing. And this module also abounds with problems and projects for the student--from easy to challenging. We all found the module interesting and novel; the students reacted very favorably to the relevancy of the material to their life. They read the module, worked many of the problems, and went to the reference material.

The list of references given is excellent. Unfortunately some of the material referenced was difficult to impossible for us to obtain, particularly papers in some law journals and RAND reports.

One day was set aside for a lecture on Stirling's formula and the gamma function. This was one of the few formal lectures given in the course up to this point. (Recall this was the third module studied.)

The students probably liked this module best of the four we studied.

I knew from examining the Heat Transfer in Frozen Soil module that it required more mathematical background than my students had. However, I wanted them to see a good engineering application. Only one of them had been exposed to engineering or standard applied mathematics courses. So I decided to do as much of this module as possible, filling in background where needed with formal lectures and outside reading assignments. Thus the teaching method changed at this point to the more traditional one of lectures and problem assignments, although discussion, questions, and diversions were still encouraged.

To someone contemplating using the modules I recommend that they obtain them a semester in advance in order to go over them in detail carefully observing what prerequisites will be required, obtaining needed reference material, and composing auxiliary material such as exercises, projects, and teaching aids. I plan to teach such a course again next year and am going to strive to have a modeling course made part of our mathematics curriculum. To quote one of my students concerning the course, "It is a great finish to a four-year mathematics program."

5. Roger Pitasky, Marietta College

Marietta College is a small liberal arts college of 1700 students. There is no engineering school; there are 200 students majoring in economics and business out of only 800 who have declared majors.

Most of the mathematics majors are not what we would have thought of as mathematics majors a few years ago. Now, probability and statistics and linear programming are in vogue and things like classical applied mathematics, numerical analysis and advanced calculus are not.

For many years Marietta College offered a two-semester course in Applied Mathematics which was strictly physical science oriented. Indeed until recently it was a joint offering of the departments of physics and mathematics. Even recently, when the course was taught solely by the department of mathematics, the bulk of the students were physics majors (who are required to take at least the first semester), along with a few mathematics majors and an occasional major in petroleum science or chemistry. Lately, we were attracting only 6-8 students for the first semester and 3-5 for the second. In the spring of 1973 our department discussed the problem of how to enlarge the audience. It was suggested that we split the course into two autonomous halves. The first semester would remain the way it was, but the second semester would be a course in mathematics applied to non-physical sciences. I was selected to be the initial instructor of this as yet unformulated course. It was only later that the connection with the CUPM project was made.

I recruited the students on the basis that

(1) the course would be experimental

(2) we were going to study realistic problems

(3) we were going to study a small number of examples but these would be considered from beginning to end

(4) we would concentrate the bulk of our time on "soft science" applications.

The result was a class of 18 students--very large by historical standards for this course. The composition of the class is worth noting too. Of the 18 students, 10 were mathematics majors, 4 others were mathematics majors with special interests in business or economics, and only 4 were physical science majors. Almost half the class had a limited knowledge of physics, and some students had never had a course in differential equations. Only six had taken the first semester of applied mathematics.

The course was taught in a traditional lecture style. In particular, the computer did not play a central role in the teaching method.

By most criteria, the experiment was a definite success:

a. From the standpoint of the goals of the experiment. In the proposal for this experiment, two goals emerged: (1) to create an Applied Mathematics course in which realistic problems could be studied, especially from beginning to end, and (2) to create a course in which model-building was a central theme. Our section did not fully succeed in studying real-life problems, nor did we succeed in studying all aspects of the problems we did study. Thus goal (1) was not fully attained. However goal (2) was realized, so well in fact that we are considering making our course into one in model-building.

b. From the standpoint of suitable materials. We covered background material on modeling, Voting Systems, Ecosystems and Population Mathematics. When I teach this course again, I plan to repeat at least two of these.

c. From the standpoint of creating a new, interesting course. Marietta College was interested in changing one semester of its Applied Mathematics sequence to a more viable course. A year ago, we had real doubts as to whether we could succeed. Now we have great confidence that we can. I believe the real success of this experiment will show up not in replacing current Applied Mathematics courses, but in augmenting them with courses using the new approach.

I have come to the conclusion that perhaps realistic problems were too much to hope for in the first place, at least at four-year liberal arts colleges. It would certainly take the right teacher with the right class to have it work. Perhaps a course in model-building with one good semi-realistic example would be more attainable and just as valuable. While teaching this course, I was constantly being reminded that there is so much basic mathematics which must be understood thoroughly before the student can appreciate some of the specialized tools that realistic problems expose them to. Despite my initial enthusiasm for bringing the students right to the brink of the action, I now wonder if we are being the counterpart of the music teacher who tries to teach a concert piece to the student before he has completely mastered the scales. The student might memorize the score, but the technique will be lacking.

The student reaction to the course was very favorable: our department chairman conducted a survey of our students in which, among other things, he was interested in which courses they particularly liked. A disproportionate number of students liked the experimental Applied Mathematics course; found it especially interesting.

It is especially impressive that the experiment succeeded in light of some of the obstacles it faced, obstacles, at least, to an organized lecture style presentation along traditional lines. For example

a. There were no individual copies of the modules for the students. This put quite a bit of added pressure on the lectures since every example and every homework problem had to be given directly to the class. This slowed down the pace of the course, doubly penalized the student who missed a class, added to the bad effects of any errors committed in class, and deprived the student of some background material.

b. The material was less familiar to me than the traditional Applied Mathematics subject matter and I lacked adequate opportunity for advance preparation. In a couple of cases, I knew very little beyond what I taught to the class.

c. The modules themselves had certain deficiencies. In most cases they were first drafts.

d. There was only six weeks between the time I got the modules and the beginning of classes. After I began the course by covering chapters I and II of Maki and Thompson, the only module available in what seemed to be near its final form was Lucas' on voting power. A week before I finished that module, a revision of "Ecosystems" arrived, so I covered it next. It is a bit ironic that my teaching style happens to include a great deal of organization--I often hand out a syllabus the first day of class--and here I was making decisions based on yesterday's mail (or lack of it).

e. I was unable to do everything I planned to do. At one point I had high hopes of including some simulations. I just never found the time to write the required programs. Judging from the bibliographies I got, I should have spent a month in the college library and should have had a sizable set of materials on reserve. I couldn't begin to find the time. The course undoubtedly suffered from the lack of the programs and collateral readings. Yet even without these activities, I clearly spent much more time on this course than a typical one.

It appears that the idea of a course in model-building is so sound that it overcame a partial lack of suitable materials, adequate course planning, and expertise of the instructor--impressive indeed.

6. Kenneth R. Rebman, California State University, Hayward

A course utilizing three of the modules was offered during the Spring Quarter, 1974. In order to facilitate scheduling, we adopted the name and number of an existing course: Math 3800 - Topics in Applied Mathematics. The catalog description of Math 3800 details a fairly standard undergraduate course in mathematics applicable to the physical sciences. In order that students not be misled as to what was actually to be offered, I wrote a one-page notice that was given wide circulation. Several colleagues were quite intrigued by the possibilities of this course, and encouraged some of their students to participate.

There were initially 14 students enrolled. During the quarter, two students dropped, primarily because of outside work pressures (not uncommon at Hayward). Of the remaining 12 students, 11 were senior mathematics majors, and one was a senior physics major. Their academic averages ranged from C+/B- to A. All had a course in differential equations and most had a linear algebra course in their background.

At the first class meeting, I explained that we would be behaving, as much as possible, like applied mathematicians. That is, our central concern would be the examination of a particular problem; the mathematics that we learned along the way would be that which was relevant to our problem-solving. Indeed, it is this feature that seems to distinguish this project from other applied mathematics materials recently developed.

I then sketched, very briefly, the nature of the three problems we were to examine. I told them that the elements of each problem would be covered in some detail during the class, but that further specific work was to be done independently by smaller groups, one to a problem. Thus I wanted them to be at least vaguely familiar with all three problems at the outset, so they could make an intelligent choice of groups.

We used Weighted Voting Systems, Two-Species Ecosystems, and Frequency Response Methods.

Weighted Voting Systems (Lucas). For this module, we spent most of our time utilizing an "open-ended" discussion. In general, I tried to instigate certain lines of inquiry, and hope that it would be followed up by the students. For the most part this was quite successful. The students enjoyed this departure from the usual classroom lecture technique and several (but not all) were eager to participate in the general discussion.

Throughout the course, we would take appropriate excursions into side areas of mathematics that were relevant to the problem at hand. In this module, there were three particular topics that arose naturally, and provided the basis for an hour's lecture/discussion. The students seemed a bit unused to the idea of not having formal lectures about mathematics that was new to them. Hence, these mini-topics seemed to satisfy their need for some standard classes.

All during the course of this module, I assigned homework problems of various types. These were assigned to all students, not just the one group. Several of these involved hand calculations of both the Shapley-Shubik and Banzaf indices.

The general classroom discussion of this module continued for about four weeks (12 meetings). At the end of that time, five students indicated their interest in doing a project in this area. Eventaully they did much more on their own initiative than I would have required.

This module was extremely successful. The entire class was interested in the general discussion; and the small group was highly motivated to tackle a major project.

Two-Species Ecosystems (van der Vaart). Initially, this module was begun using the same informal discussion technique that had worked so well for the voting module. However, it soon (after two or three class meetings) became apparent that this would not be very successful. Although I continued to maintain a relaxed and informal atmosphere, I found it necessary to give essentially standard lectures. I suspect there were several underlying reasons causing this change to a more traditional approach:

a) The mathematics required was of a more advanced nature than that necessary in the previous module. And, although the students all had experience with differential equations, they had virtually no familiarity with the qualitative study of systems of differential equations.

b) The idea of applying mathematics in a biological model was quite new to most of them. All of us were self-conscious about our lack of biological expertise.

c) I myself was much less familiar with the background information (both mathematical and biological) used in this module than I was with the background material in the voting module. I think it is considerably easier to give a lecture than to preside over an open-ended discussion. Thus, while I could prepare good lectures, I really wasn't competent to direct a worthwhile discussion session.

The general mathematical theme throughout our study of this module was the qualitative theory of systems of differential equations. As in the voting module, I had some clear-cut opportunities to offer a few standard lectures on mathematical topics required for our study of the particular problem. We did an extensive review of linear algebra, particularly the theory of eigenvalues. We then saw an application of eigenvalues in examining the nature of critical points of 2 × 2 systems of linear first-order differential equations. Of course, our system were non-linear, and the students could now appreciate the great value of the theorem that says that the behavior near the critical-points of non-linear systems is determined by the behavior of critical points in the linear approximation. Thus the importance of this quite difficult theorem was made abundantly clear.

The classroom discussion of this module continued for about four weeks. In that time we essentially covered Sections 1 and 2. Of course I omitted some of this material and inserted several mathematical lectures of my own. Many of the author's questions were presented as homework exercises. The class was also given a copy of the bibliography, and several of them (but not those already committed to the voting problem) made use of this.

At the end of the third week of our study of this module, the group that would work on this project identified themselves. This group also consisted of five students. It is now clear that I really did not leave sufficient time for this group to proceed as deeply as they would like. However, in the time available to them, I think they did an outstanding job on their project.

Frequency Response Methods (Powers). When we arrived at this module, there were but two weeks remaining in the quarter. I would have been content to let the other groups simply work out their projects. But since I had agreed to test this module, and since there were still two students without a project, we went ahead with it.

I first described the very rough outline of the problem to these two students. They immediately recognized the model as that of a damped harmonic oscillator. (One student was a good math major with a strong background in differential equations; the other was the physics major.) Although they were not interested in doing the experiment described in the module, the intrinsic idea of actually using an experiment to find the underlying differential equation was extremely interesting to them.

I wanted to test the author's claim that students would be able to acquire the necessary mathematical information about systems analysis, transfer functions, Bode plots, etc., on their own. So after our initial discussion of the problem I gave them the list of references and told them to return in a few days. They were extremely successful in obtaining most of the necessary information in a relatively short time. I had to fill in a few gaps, but this was primarily in the interest of saving time.

Once they had the necessary background, we turned to the actual problem. As a substitute for the suggested experiment they devised what I believe is a rather ingenious approach. They programmed an analog computer with a second order linear differential equation and plotted various input functions $u(t)$ and the resulting output functions $y(t)$. These plots were quite accurate; varying the frequency of the input, they could measure the amplitude and phase shift of the steady-state output. They generated approximately 30 pairs of such graphs. It was then their intention to continue the experiment as described, but using the data they had collected from the analog plots. In this way, they hoped to be able to recover the original equation that had been programmed into the analog computer. Unfortunately, lack of time prevented them from completing this. They were quite excited about the whole project however, even to the extent of wanting to finish during the summer.

I was struck by the virtually unanimous desire by the students to utilize the computer. Moreover, I was truly impressed by the quite sophisticated results they were able to obtain. They pointed out, however, that it was extremely helpful (perhaps essential) that two of them were regularly employed as operators at the campus computer center.

One impressive thing to me about this course was the willingness of some marginal students to get really involved in the subject matter. Certainly, those students with a good academic record were still the good students; but some C students worked just as hard and just as successfully. I felt no compunction at all about awarding all A's. This is, I guess, just one more manifestation of the high student interest in the course. Another side effect was a much lower than usual rate of absenteeism. Indeed, whenever several students were absent, I would usually later discover that they were meeting together on their project. Also unusual was the suggestion of some students to continue their project throughout the summer.

The student response to the class was overwhelmingly favorable. They commented that it was really interesting and exciting to them to be able to synthesize lots of different kinds of mathematics and bring it to bear on a particular problem. (I did observe a readiness on their part to set up models that were much too difficult and general. Much of the guidance of open-ended discussion was towards simplifying the models.) The one adverse criticism, which also was universal among the students, was that there was not enough time. If I were doing the course again I would only do two modules and try to run them parallel for the first few weeks, thus letting all students have more time to work on individual projects.

I think that our course was a distinct success. Both I and my students appreciated the opportunity to participate in this venture. I will be exploring the possibility of adding a models course such as this to our curriculum.

7. Kenneth Stofflet, University of Evansville

Modules Tested: Dynamics of Several Species Ecosystesm; Heat Transfer in Frozen Soil; Network Analysis of Steam Generator Flow; Population Mathematics (independently by one student).

The modules were tested in Mathematics 460, Topics in Applied Mathematics. (A special creation which plans to live on.) Enrollment, 11; credit, 2 quarter hours; prerequisite, consent of instructor. As it turned out 10 of the students had recently completed a quarter of differential equations. The 11th, a freshman, was concurrently enrolled in a differential equations course and Mathematics 460. All had some previous exposure to a computer.

The class met for two hours each week--roughly seven class hours per module. The class time was spent in discussion, student presentation and instructor presentation, with too much of the latter. After the third week the students worked outside of class in teams of one, two, or three on special projects.

Students: An abbreviated student description is given below. All of them had been in classes with me before. I would say they ranged from average to exceptional in academic ability. I thoroughly enjoyed working with them.

Year	Major	Cum. GPA	Goals
Sophomore	Math-Physics	3.80	Graduate School
Senior	Physics	3.55	Utah State Acoustical Engineer
Sophomore	Mathematics	3.35	Graduate School
Freshman	Math-Physics	3.85	Ph.D.
Sophomore	Engineering	2.81	Engineering Degree
Junior	Engineering	2.29	B.S. Engineering
Junior	Math-Physics	3.79	Applied Math Ph.D.
Senior	Mathematics	2.81	M.S. Math or Physics
Senior	Chemistry	2.34	B.S. Degree
Graduate	General Education	4.00	School System Administration
Junior	Mathematics	2.27	B.S. Degree

Reaction: Without exception the material was considered interesting and was worked at enthusiastically. I can honestly say I've never enjoyed teaching a class more. (Several equals but no greater-thans.) The only real complaint was a lack of time.

Evaluation: My evaluation was mainly subjective--I guess it always is. Class participation was good and some days we actually got rather lively. (Occasionally heated.) Each student worked on a project and talked with me about it. They convinced me they understood at least what they were doing. No one worked all the exercises--including me--but everyone worked some and roughly half were discussed.

Role of the Computer: Definitely an important part of our course. We used it early, in time sharing mode, to illustrate stability problems with some difference equations, as suggested by exercises in the Ecosystems module, and to look at some Lotka-Volterra two species models.

The culmination of each module should have been the analysis and discussion of a computer output. We weren't that good.

Observations and Guidelines.

Materials of the kind developed for this project can be taught with great success both by traditional lecture methods and by various open-ended techniques.

Regardless of the method used, the teachers reported that they had to work much more on this course than on a regular upper division course. This was due in part to the experimental nature of the classes--for example, being under some obligation to test a given number of modules (in draft form)--but there is clearly a built-in extra burden for the teacher of this course.

The extra work is compensated for, if not partially caused, by the fact that students also worked harder on the course.

Everyone agreed to the estimate that three or four modules could be done by the lecture approach, but that only two should be attempted by open-ended methods in one term.

Classes of less than twenty students were seen as almost essential for the use of open-ended methods.

Students were generally enthusiastic about the way this course drew upon a variety of their earlier mathematical studies and put them to use in an applied context.

The practice of assigning projects to groups of students worked extremely well wherever it was used.

In a number of cases, the students' interest and ability in pursuing computational results was surprising, sometimes providing a major motivation. On the other hand, reliance on computation did not seem to be necessary for a successful course. The possibility of using canned programs was mentioned, but the main enthusiasm for computing involvement was precisely due to the development and testing of programs.

When a class was presented with just some background information and asked to develop a model, the usual result was a very complicated model greatly in need of further discussion. An exception seems to have been the concept of voting power; two classes were able to arrive at the standard models.

Students were uncomfortable in their new role in open-ended mathematics classes. Occasional lecturing as an antidote to this feeling worked rather well.

Those teachers attempting the open-ended approach found themselves forced back on traditional lectures when time became short and when the material started to exceed the level of the students' backgrounds.

The material prepared by the project was written for the teacher, to be provided to the students only according to his plan. This caused extra difficulties for the teacher using traditional methods, e.g., making reading assignments. The problem was partially solved by one teacher by distributing the daily lecture notes of a "student recorder."

Since the objective of open-ended teaching would be subverted by providing the class with the module containing the refined model they are seeking, this remains a problem. Some of the revisions that were made after the conference were intended to make the modules more appropriate for direct student use. They remain, however, resource materials for the teacher primarily.

Last but not least of the observations: it is much easier and takes much less preparation to give a lecture than to conduct an open-ended class successfully. "The classes I spent the most time preparing were the ones where I said the least."

Chapter 3

MEASURING POWER IN WEIGHTED VOTING SYSTEMS

William F. Lucas
Department of Operations Research
Cornell University

PREFACE

This paper describes two numerical indices, due to Shapley and Shubik and to Banzhaf, which are useful for measuring political power in various weighted voting schemes in which some voters effectively cast more or heavier weighted votes than others. A rather detailed review is given of the many uses of these indices in practical situations, as well as some suggestions for potential new applications, research projects, and future directions. It is intended that this presentation serve both as a survey for the interested reader, and as an educational work useful in the college classroom. As a result of this dual purpose, it has also been published as Technical Report No. 227 from the Department of Operations Research at Cornell University (Ithaca, New York 14853; September 1974).

There are only a very few simple mathematical concepts among the few formal prerequisites required of the reader. Most of this paper can be followed by one with a familiarity with the most basic concepts from set theory plus a knowledge of permutations and combinations from elementary algebra. Most of the analysis is of a discrete or combinatorial nature. However, some of the arithmetical or computational techniques are fairly complicated in some parts of the later sections, and a higher level of mathematical maturity or sophistication will be required at these points. Nevertheless, some parts of the paper are suitable for use at various levels from secondary school through college and beyond. On the other hand, the material about large scale computation in Section 6 is somewhat more difficult and presumes some knowledge of more specific college mathematics; the details of which can be seen by skimming through this Section. And it is most desirable that anyone who will be involved with such calculations have some computer experience beforehand. The user should be alerted to the fact that some part of Section 6 will be necessary for anyone who wishes to go into the detailed computational projects suggested in Section 7. The instructor however should be easily able to fill in most necessary details. Sections 8 and 9 are independent of the earlier sections.

A large number of projects are recommended throughout the paper. These vary in length from routine class exercises (some with answers and some without), to mini projects which should take a few days to a couple of weeks, to major projects such as in Section 7 which should take from about four to eight weeks, and finally to some suggestions which can develop into long term original research topics. The mini projects in Section 5 are often appropriate for student teams of two or three persons, and those in Section 7 could involve up to about six people. The instructor may decide to give none, some, or all of this material to the students as best fits his needs. There may be some advantages in withholding much of this paper in the early stages so that the students can get more practice in the actual model building aspect of research. They may, after examining several examples, be able to derive one of these two power indices, or an alternate one, by themselves. This more undirected approach will of course lead to student doubts and anxieties, and to crises in confidence; and the instructor will continually need to provide guidance and direction, and to rekindle enthusiasm by feeding in bits of new information. It is also recommended that the instructor or students make some contact with colleagues in their school's political science or government department. Such "outsiders" may serve in various capacities from mere critics up to members of a team teaching or research endeavor.

The intention was to write this paper so as to be fairly self-contained and to require few, if any, of the works listed in the references. Some of the unpublished reports, RAND memoranda, and papers appearing in law journals might not be readily available to all instructors, and it may take some lead time to acquire them. It will be fairly clear in reading Section 5 as to which references should be available in order to do particular mini projects. The paper by Banzhaf (1966) or Johnson (1969) should prove most useful in the latter parts of Section 5. Banzhaf (1968), Peirce (1968), and perhaps an appropriate reference mentioned in Section 6, would be helpful for doing the projects in Section 7. The books by Farquharson (1969), Papayanopoulos (1973), Riker (1962a), and Shubik (1964), as well as those containing the papers by Riker and Shapley (1968) and Shapley (1953) should be available in most school libraries, and can serve as more general reading. Also, see the recent book by Brams (1975).

An incomplete first draft of this paper, dated October 1973, was used in one way or another by about a dozen teachers, including the author, in the spring of 1974. A few of these instructors were part of the official CUPM evaluation, but most of them had contact with the author directly. Preliminary reports indicate a fair amount of success. Incidentally, in at least three of these classes the students arrived at one of these two power indices by themselves--with some hints perhaps. The author is grateful for the comments from these instructors, from the other members of this CUPM panel, as well as to Louis Billera, Pradeep Dubey, Irwin Mann, Joseph Malkevitch, Lloyd Shapley, and Robert Weber who read and made suggestions on the draft version. Hopefully, the number of remaining errors in this revision and extension are rather small; but then again it is not an uncommon part of the model building process to detect mistakes in previous studies.

1. INTRODUCTION

1.1. Weighted Voting. There are a large number of voting situations in which some individuals or blocs of voters effectively cast more ballots than others. Such weighted voting systems are found in governmental bodies such as the U. S. Congress, some state legislatures and county boards, in the Electoral College, in voting by stockholders in a corporation, in several university senates, in many other multimember electoral districts in which several representatives are elected at-large from a single district, as well as when strictly disciplined political parties vote as a single bloc.

A prime concern in designing such policy or decision making assemblies usually is that they are fair and equitable in the sense that they give equal representation to their constituents. There have been many challenges in recent times against existing legislative bodies. There are frequently heard charges of undemocratic inequity, unfairness, disparities, bias, handicaps, debasement, denying, gross variances, impairment, prohibitions, expediency, malapportionment, disenfranchising, ambiguous and archaic systems, diluting and devaluating the vote, and "stuffing" the ballot boxes. The courts have frequently found deliberative and legislative bodies wanting, and have often ruled in favor of impartiality, equal protection, political justice, full participation and representation, the right of suffrage, and the principle of "one man-one vote" for all citizens.

1.2. Voting Power and Its Relation to Weights. Some proponents of weighted voting argue that such systems can be used to adjust for or to cancel out inequalities at another level, such as differences in the populations which various delegates represent. However, the fraction of the total number of votes which a representative possesses, or the number of members in a party, is not generally synonymous with any meaningful measure of his or its voting power! The ability to cast more ballots does not in itself necessarily increase one's power nor does it do so in a directly proportional way. There may exist major discrepancies between the weights voters have and a good measure of their influences. It is fallacious to expect that one's voting power is directly proportional to the number of votes he can deliver. Yet many attempts to correct inequalities merely assign weights to a delegate proportional to the number of inhabitants he represents, and it is felt that this preserves some equality at the level of the individual citizens. Paradoxically, those who advocate that they are the main beneficiaries of the weighted systems such as the Electoral College are very often the ones most hurt by it in terms of power indices discussed below. The discrepancies between the ratio of weights and the ratio of power will be illustrated by several examples in the following sections. Power is not a trivial function of one's strength as measured by his number of votes. Simple additive or division arguments are not sufficient, but more complicated relations are necessary to understand the real distribution of influence.

1.3. Applications. The power indices discussed below do not constitute an attack on all weighted voting systems per se. Instead, they provide a method for structuring such systems with more understanding, so that certain individual powers can be preserved. Presumably the courts are against indirect or sophisticated discrimination as well as the more obvious types. Weighted voting with periodic adjusting of the weights, when done properly, can perhaps serve as an alternative to creating other inequities, caused by events such as districts of varying size due to shifting populations. It can be used to maintain present advantages or to avoid going through frequent redistricting, reapportionment, realignments, or otherwise disturbing existing constituencies, such as cutting county boundaries to obtain state wide districts; which in turn may give rise to other potential difficulties such as gerrymandering. Proper weighting can possibly compensate or offset for some other problems or defects which arise in attempting to make all districts "equal."

1.4. Measures for Voting Power. Several examples of weighted voting systems will be scrutinized in what follows in order to gain some insight into the complexities and ramifications inherent in the apportionment of unequal voting strengths. Two of the numerical indices which have been used to measure a subject's share of the power to control or influence outcomes will be discussed. This pinpointing of some of the more formal, technical, and quantifiable aspects of bloc voting should clarify many of the implications and consequences which are built into such divisions of power. From a more complete account and accurate evaluation, one should be able to conclude in a more intelligent manner where the relative power to influence and affect aggregate decisions or laws resides.

These analytical techniques should prove useful in re-evaluating some of the existing democratic institutions in terms of fairness, concealed biases, and the degree of protection they provide their constituents. Procedures for adequately modifying, altering or revising current assemblies to reflect certain changes should be less ambiguous. These methods can be used to evaluate a newly proposed system, or to design a new legislative body from scratch like a college senate (either a hypothetical or proposed one, or to revise an existing one) so as to provide each with the influence he is entitled to. Of course, no scientific theory or mathematical formula will suffice to reveal all of the poorly understood and subtle details, nor the behind-the-scene nuances, which can have significant influence on the outcomes reached by voting in a deliberating or legislative council. Additional comments on the possible inadequacy of such theories are contained in Section 9.

1.5. Outline. Some simple examples of weighted voting systems plus some notation and definitions will be presented in Section 2, and some less trivial examples will be mentioned and evaluated for power ratios in Section 4. The power indices introduced by Shapley and Shubik (1954) and by Banzhaf (1965) will be discussed in Section 3. Some mini projects are given in Section 5, and these may be undertaken with or without the suggestions for large size calculations and machine computation described in Section 6. The main project of evaluating the current Electoral College and some of its proposed alternatives is presented in Section 7. Some possible extensions and generalizations for this study and other work on voting appear in Section 8, and some criticisms of this whole approach are mentioned in the final Section.

2. SIMPLE EXAMPLES

2.0. Introduction. Several simple examples of weighted voting schemes will be considered in this section. We will first introduce the notation for a weighted majority game:

$$[q;w_1,w_2,\ldots,w_n].$$

Here there are n voters, or players or citizens, denoted by $1,2,\ldots,n$; and w_i is the voting weight of player i and is assumed to be a nonnegative number. Let

$$N = \{1,2,\ldots,n\}$$

be the set of all n voters, and let S be a typical coalition or bloc of voters, i.e., a subset of N. A coalition wins a vote, or is called winning, whenever

$$\sum_{i\in S} w_i \geqq q$$

where we will assume that

$$q > \tfrac{1}{2} \sum_{i\in N} w_i .$$

This q is called the quota for the game. In some cases, but not all, one also assumes that each $w_i < q$, and one often restricts the w_i and q to integer values.

2.1. Stockholders. The per cent of a company's stock which an investor owns can serve as a weight in a weighted majority game, but it is often not a good measure of his influence or voting power. A man or coalition with over 50% of the stock has full control or has 100% of the voting power whenever decisions are made by a simple majority, i.e., is a dictator. A group with exactly 50%, or more, of the stock can veto or block any such action. In the game [51;28,24,24,24] the first voter seems to be much stronger than the last three, since he needs only one other to pass an issue whereas the other three must all combine in order to win. In the situation [51;26,26,26,22] the last player seems powerless since any winning coalition containing him can just as well win without him. The two games [51;40,30,20,10] and [51;30,25,25,20] seem identical in terms of voting influence, since the same coalitions are winning in both cases.

2.2. Equal Power. It is clear that in the game [q;1,1,...,1] each player has equal power. Such games arise in pure bargaining situations, and in deterrence encounters where each participant has the same amount to lose. However, games such as [3;2,2,1], [8;7,5,3] and [51;49,48,3] are similar to [2;1,1,1] in terms of power, since they give rise to the same collection of winning coalitions. The last of these illustrates the potential value of a small third party such as the Liberal Party in Great Britain. If we add to the game [3;2,1,1] the rule that player 2 (or 3) can cast an additional vote in the case of a 2 to 2 tie, then it is effectively [3;2,2,1]. But if player 1 can cast the tie breaker, then it becomes [3;3,1,1] and he is a dictator. More generally, in the game [50(n-1)+1;100,100,...,100,1] player n has the same power as the others when n is odd, i.e., the game is similar to one in which all of the players have the same weights or number of votes.

2.3. Dummy Players. A player who has no real voting power is called a dummy. Any winning coalition which contains such an impotent voter could win just as well without him. Player 4 in the game [51;26,26,26,22] is an example. Also player n in the game [50(n-1)+1;100,100,...,100,1] is a dummy when n is even. In the game [16;12,6,6,4,3] player 5 with 3 votes, is a dummy; since no subset of the numbers 12,6,6,4 sums to 13, 14 or 15, and thus player 5 could never be pivotal in the sense that by adding his vote a coalition would just reach or surpass the quota of 16. On the other hand, one can show that player 4 with 4 votes has the same influence in terms of winning coalitions as either player 2 or 3 with 6 votes apiece. This example can be generalized to games like [268;48,36,30,...,6,4,3] where the missing weights are multiples of six. The last three players in the situation [10;5,5,5,2,1,1] are all dummies.

Banzhaf (1965) describes a weighted voting system which was used in Nassau County, New York by the Board of Supervisors in which three of the six municipalities share all of the power equally and three others are dummies in both instances, and this is described in Table 2.3.1. Each municipality has one representative whose vote is weighted according to the number of votes listed in the Table. Of course, the dummies may be able to influence or sway decisions by their right to debate or in some other way, such as by membership in committees. A more recent revision of this Nassau County example which avoids dummies is mentioned in Section 5.6.

NASSAU COUNTY

Municipality	Population 1954	No. of Votes 1958	Population 1960	No. of Votes 1964
Hempstead, No. 1	618,065	9	728,625	31
Hempstead, No. 2		9		31
North Hempstead	184,060	7	213,225	21
Oyster Bay	164,716	3	285,545	28
Glen Cove	19,296	1	22,752	2
Long Beach	17,999	1	25,654	2
Total	1,004,136	30	1,275,801	115

TABLE 2.3.1

The Canadian election in the fall of 1972 gave the results in Table 2.3.2, where each of the three listed parties had equal power and those in the "others" category constitute a dummy (individually or as a group). The Liberals have since won a majority in the elections of 1974.

CANADA

Party	Leader	No. of Seats
Liberals	Pierre E. Trudeau	109
Tories	Robert L. Stanfield	107
New Democrats	David Lewis	31
Others		17

TABLE 2.3.2

2.4. Veto Power and Dictators. A player or coalition is said to have veto power if no coalition is able to win a ballot without his or their consent. A voter is a dictator if he controls any vote by himself, i.e., his weight $w_i \geqq q$. There can only be one dictator (recall that we assumed

$2q > \Sigma w_i$), and the other players are then all dummies. Player 1 has veto power in the game [51;50,49,1] which is essentially the same as [3;2,1,1]; and if he is a chairman with additional power to break ties then the game becomes [3;3,1,1] and he is a dictator. Note that the ability of an individual to break tie votes in [[n/2]+1;1,1,...,1] adds power when n is even and adds nothing when n is odd. This presumes no change in q due to absences or abstentions. Here, [n/2] stands for the greatest integer in n/2.

3. MEASURES OF VOTING POWER

3.0. Introduction. Several quantitative measures for evaluating the abstract power of a voter or coalition have been proposed. The Shapley-Shubik index as well as the Banzhaf index will be introduced in Sections 3.3 and 3.5, respectively. These two value concepts have received the most theoretical attention as well as application to real-world political structures. These indices will be computed for some very simple examples throughout Sections 3 and 4, and applied to some more realistic situations in Sections 5 and 7.

3.1. The Notion of Power. It is clear from our examples that voting power, expressed in some formal or functional sense, is not directly proportional to the number of votes one casts. In fact, an exact proportionality between weights and the power indices defined below seems to be rare in cases other than the one man-one vote situation. We are thus concerned with developing indices which give a reasonable measure of voting power in some technical sense. An individual's index should relate, in a preferably simple way, to one's true ability to affect group decisions by means of casting his weighted vote; and it should hopefully capture and crystallize some of our more intuitive concepts about power, such as why some legislators seem to have a better chance on average of being on the winning side than another. It should indicate one's relative influence, in some numerical way, to bring about the passage or defeat of some bill. It should be based somehow upon the importance of the individual in casting the deciding vote which will guarantee that some issue will carry. It should compare all the opportunities which each voter has to be a sort of critical swing-man in causing a desired outcome. This index should depend upon the number n of players involved, on one's fraction of the total weight, and upon how the remainder of the weight is distributed.

Power depends essentially upon being on the winning side of a division, that is, upon the ability to succeed in joining a winning coalition which supports one's views. Furthermore, it is the possibility of membership in the "minimal" winning coalitions which is sufficient or significant; since any additional votes afterwards become unnecessary and irrelevant, whereas deleting a member beforehand fails to insure a victory. It is what one brings to a successful alliance that is crucial, and the one who is "last" to join a minimal winning coalition is particularly influential. The possibility of reversing an outcome by changing one's vote on a question is a most important one. The person whose support is necessary for success has a rather strong bargaining position, and one in this uniquely influential marginal or pivotal position is often rewarded handsomely for casting the deciding vote which carries his coalition "over the top." It is assumed for an abstract theory that all theoretically possible voting alignments, or combinations, are equally likely to occur, and thus the scientific measures which are derived below ignore a large number of less technical concepts such as voting in alphabetical order, political ideologies, and many other aspects as will be mentioned in Section 9.

3.2. Terminology. As we mentioned above, a coalition of players which possess enough votes to guarantee passage of an issue is called winning. A winning coalition is said to be minimal winning if no proper subset of it is winning, and these are the important ones in the following Section.

A coalition which is not winning is called losing. A subset S of voters is a blocking coalition or has veto power if its complement N-S is not winning. (Some authors also require that S itself be losing in order to be a blocking coalition.) A player i is a dictator if he forms a winning coalition {i} by himself. A voter i is a dummy if every winning coalition which contains him is also winning without him, that is, he is in no minimal winning coalition.

There are many interrelations and properties for the concepts just defined; especially if additional conditions are assumed, such as our quota q consisting of a simple majority of the total voting weights. In fact, a whole algebra or combinatorial structure of simple games can be built up from these ideas as illustrated in Shapley (1962) and by later generalizations, but this will not be done here. The interested reader should see page 87 of Shapley (1973) for references on this and some related theories. A good exercise at this point would be to describe which coalitions in the examples of Section 2 are of the various types mentioned in this Section. It may prove helpful to actually draw the lattice of all subsets of the set of players for these examples.

3.3. The Shapley-Shubik Index. Shapley and Shubik (1954) introduced an index for measuring an individual's voting power, which is a special application of a more general value concept introduced by Shapley (1953) in the context of the general von Neumann-Morgenstern (1944) theory of multiperson cooperative games. A voter's value is the a priori chance that he will be the last member added to turn a losing coalition into a winning one. It thus assigns him the probability of his casting the deciding vote. This expected frequency with which a man is the pivot, over all possible alignments of the voters, appears to be a good indication of his voting power.

More precisely, one looks at all possible orderings of the n players, and considers this as all of the potential ways of building up toward a winning coalition. There are $n!$ such ordered sequences, or permutations, of the n voters. The "!" stands for factorial; $n! = n\times(n-1)\times\ldots\times 2\times 1$ and $0! = 1$. For each one of these permutations, some unique player joins and thereby turns a losing coalition into a winning one, and this voter is called the pivot. That is, if in the sequence of players $x_1, x_2, \ldots, x_{i-1}, x_i, \ldots, x_n$, $\{x_1, x_2, \ldots, x_{i-1}, x_i\}$ is a winning coalition, but $\{x_1, x_2, \ldots, x_{i-1}\}$ is losing, then i is in the pivotal position. Consider the number of permutations of the n voters in which a particular player i is the pivot; and then this number divided by the total number of alignments, which is $n!$, is the Shapley-Shubik power index or voting value, and is denoted by φ_i:

$$\varphi_i = \text{Number of sequences in which } i \text{ is a pivot} \div n!.$$

Also, let

$$\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n).$$

We are assuming here that each of the $n!$ alignments is equiprobable. One may consider the random order as a ranking of the voters according to their degree of enthusiasm or intensity of feeling in support of the issue being voted on, and the issue as a random variable. We are not concerned here with the most likely order of voting in some real convention or assembly, such as voting as you enter through some door, or alphabetically, or some other historical or traditional way; nor with developing a realistic theory of how voting coalitions do grow in practical circumstances, even though the formation of a good theory along these lines would be most desirable.

This power index can be expressed by the formula

$$\varphi_i = \sum \frac{(s-1)!(n-s)!}{n!}$$

where

$$s = |S| = \text{Number of voters in } S,$$

and where the summation is taken over all winning coalitions S for which $S-\{i\}$ is losing. This follows from the fact that there are $(s-1)!$ orders in which the given $s-1$ players can enter S before i and there are $(n-s)!$ different orders in which the remaining $n-s$ players can be added to S to form the grand coalition N.

In addition to the probabilistic approach taken above, this same index can be derived by several other approaches: by means of a somewhat dual theory using blocking coalitions instead of winning ones; from a bargaining or fair division scheme suggested by John Harsanyi (1959, 1963); or from a simple set of axioms (symmetry, additivity, and dummies are powerless) given by Shapley (1953) for his more general value concept. Shapley's axioms give a unique value formula in the more general game theory context, but not when restricted to the class of majority voting games. A variation of his axioms due to Dubey (1974) gives a uniqueness result in the case of all "superadditive simple" or "monotone simple" games. In addition to these other approaches, this index has many other nice properties, but we only mention here that the φ_i are nonnegative,

$$\sum_{i \in N} \varphi_i = 1$$

and

$$\sum \frac{(s-1)!(n-s)!}{n!} = \frac{1}{s}$$

where the summation is taken over all coalitions S with a fixed number s of players. The reader can verify these.

3.4. An Example. The 24 permutations of the four players 1, 2, 3, 4 in the weighted majority game [51;40,30,20,10] are listed below. The "*" indicates which player is pivotal in each alignment.

1 2*3 4
1 2*4 3
1 3*2 4
1 3*4 2
1 4 2*3
1 4 3*2
2 1*3 4
2 1*4 3
2 3 1*4
2 3 4*1
2 4 1*3
2 4 3*1
3 1*2 4
3 1*4 2
3 2 1*4
3 2 4*1
3 4 1*2
3 4 2*1
4 1 2*3
4 1 3*2
4 2 1*3
4 2 3*1
4 3 1*2
4 3 2*1

It is clear from this array that

$$\varphi = (10,6,6,2)/24.$$

Or, from the formula we see, for example, that

$$\varphi_1 = 3\cdot\frac{(3-1)!(4-3)!}{4!} + 2\cdot\frac{(2-1)!(4-2)!}{4!}$$

since player 1 is pivotal in 3 coalitions consisting of three players and in 2 coalitions of two players.

3.5. The Banzhaf Index. Another value concept for measuring voting power was introduced by Banzhaf (1965). He is a lawyer and much of his work has appeared in law journals; and his index, even more so than the one above, has been used in arguments in various legal proceedings. Some of the mathematical experts in this area have presented affidavits in cases before the courts (see Banzhaf (1968), page 306). Banzhaf's index is also concerned with the fraction of possibilities in which a voter is in the crucial position of being able to change an outcome by switching his vote, that is, being able to alter the whole group's decision by he alone changing. However, he considers all significant combinations of "yes" and "no" votes, rather than permutations of the players as in the Shapley-Shubik case.

Each voter can vote yea or nay on a particular question; so one can imagine all 2^n possible combinations of such votes by these n players. A player is said to be marginal, or a swing or critical, in a given combination if he can change the outcome, resulting from this combination, from passage to defeat or vice versa by changing just his vote on the issue. Power appears to rest in precisely such marginal situations in which a defection from a bare majority produces a different result, or achieves the exactly opposite goal. How often one appears in such a marginal or swing position is taken as the relative index of his influence. The ratio of voting power (in the sense of Banzhaf) for player i to that for player j is the number of combinations in which i is marginal, divided by the number in which j is marginal. Consider the number b_i of voting combinations in which voter i is marginal. The Banzhaf power index or value for player i is defined as b_i divided by the sum of all of the b_j taken over the n players. We will denote player i's Banzhaf value by β_i and let

$$\beta = (\beta_1, \beta_2, \ldots, \beta_n).$$

We again assume that all voting combinations, i.e., divisions into yeas and nays, are equally probably; since one can hardly assert in advance in the case of an abstract theory which possibilities are more likely to occur or prove most significant. We assume that everyone votes on each issue, and that no abstentions are allowed.

Note that the Shapley-Shubik index uses permutations of the players, and is concerned with the order in which winning coalitions are constructed. It assigns importance to the voter who is the last to join a coalition and thus make it a winning one. Whereas the Banzhaf index employs combinations, and considers just the number of such in which one plays a significant role. It does not look at the chronological order in which the winning coalitions were formed. It is concerned with which subsets voted yea or nay, however, and not with the numerical outcome of the vote. Comparing the values φ and β, Riker and Shapley (1968) state on page 204 that "Although the numerical values obtained are slightly different, we know of no significant qualitative differences that would arise from defining the power index in this way." Some more recent work however does suggest some differences, and with the recent axiomatization of β by Dubey in Dubey and Shapley (1975), these differences should become clear before long. Both value concepts seem to be effective and objective quantitative measures of power, which can be accurately and readily calculated in given cases. They have been accepted by several political scientists and mathematicians as reasonable and believable for understanding the logical structure of many voting problems, and they are not based on any excessive assumption.

Furthermore, both of these value concepts clearly satisfy some of our intuitive feeling towards power expressed earlier; for example, the indices are 1 for dictators and 0 for dummies, and they depend upon winning coalitions and only indirectly upon a voter's numerical weight. By definition the β_i are nonnegative and

$$\sum_{i \in N} \beta_i = 1.$$

The indices φ and β are symmetric and monotone in the sense that $w_i = w_j$ implies $\varphi_i = \varphi_j$ and $\beta_i = \beta_j$, and $w_i > w_j$ implies $\varphi_i \geqq \varphi_j$ and $\beta_i \geqq \beta_j$. One possible approach to deriving a measure of power could be to list desirable conditions like these, and to then search for a function with these properties.

3.6. Computation and Examples. To calculate β for reasonably small examples one can construct a table which lists the n players on top and the 2^n combinations of "yes" and "no" votes below the players, i.e., all the ways a vote could turn out. Next to each combination one can list whether an issue would pass or fail. Ties are not allowed. A second table aside the first one, also headed by the players, can then be used to record for each player next to each combination whether he was marginal or not. One adds up the number of times a player is marginal and the resulting numbers are proportional to the indices β_i. One could eliminate this second table of course by instead underlining or placing a "*" by the marginal voters in the first table.

The example $[51;40,30,20,10]$ considered above gives rise to Table 3.6.1. The resulting indices are $\beta = (10,6,6,2)/24$ which agree in this particular case with φ computed above. In the appendix of Banzhaf (1965) there are similar tables for the five-person games $[9;5,5,3,3,1]$ and $[5;4,2,1,1,1]$, and the results are $7\beta = (2,2,1,1,1)$ and $11\beta = (7,1,1,1,1)$, respectively. One should calculate φ for these two cases and compare it with β. A similar exercise is to compute β for $[5;4,2,1,1]$ and compare it with $\varphi = (9,1,1,1)/12$. In practice, various symmetries allows one to compute β without having to consider the full list of all combinations.

COMPUTATION OF THE BANZHAF INDEX

Players				Pass/Fail		Marginal			
1	2	3	4	P	F	1	2	3	4
Y	Y	Y	Y	P					
Y	Y	Y	N	P		X			
Y	Y	N	Y	P		X	X		
Y	N	Y	Y	P		X		X	
N	Y	Y	Y	P			X	X	X
Y	Y	N	N	P		X	X		
Y	N	Y	N	P		X		X	
N	Y	Y	N		F	X			X
Y	N	N	Y		F		X	X	
N	Y	N	Y		F	X		X	
N	N	Y	Y		F	X	X		
Y	N	N	N		F		X	X	
N	Y	N	N		F	X			
N	N	Y	N		F	X			
N	N	N	Y		F				
N	N	N	N		F				

$24 \times \beta = (10, \ 6, \ 6, \ 2)$

TABLE 3.6.1

A lattice or geometrical view for the coalitions and swings of an arbitrary four-person voting game can be obtained from Figure 3.6.1, which pictures a four-dimensional cube. Coalitions correspond to vertices. Swings for a given player correspond to certain of the edges in the direction of this particular player's dimension, as illustrated by the lines below the hypercube. This Figure illustrates swings by means of arrows in the case [51;40,30,20,10]. To obtain the number of pivots for computing the Shapley-Shubik index φ, one usually needs to count some of the swings more than once.

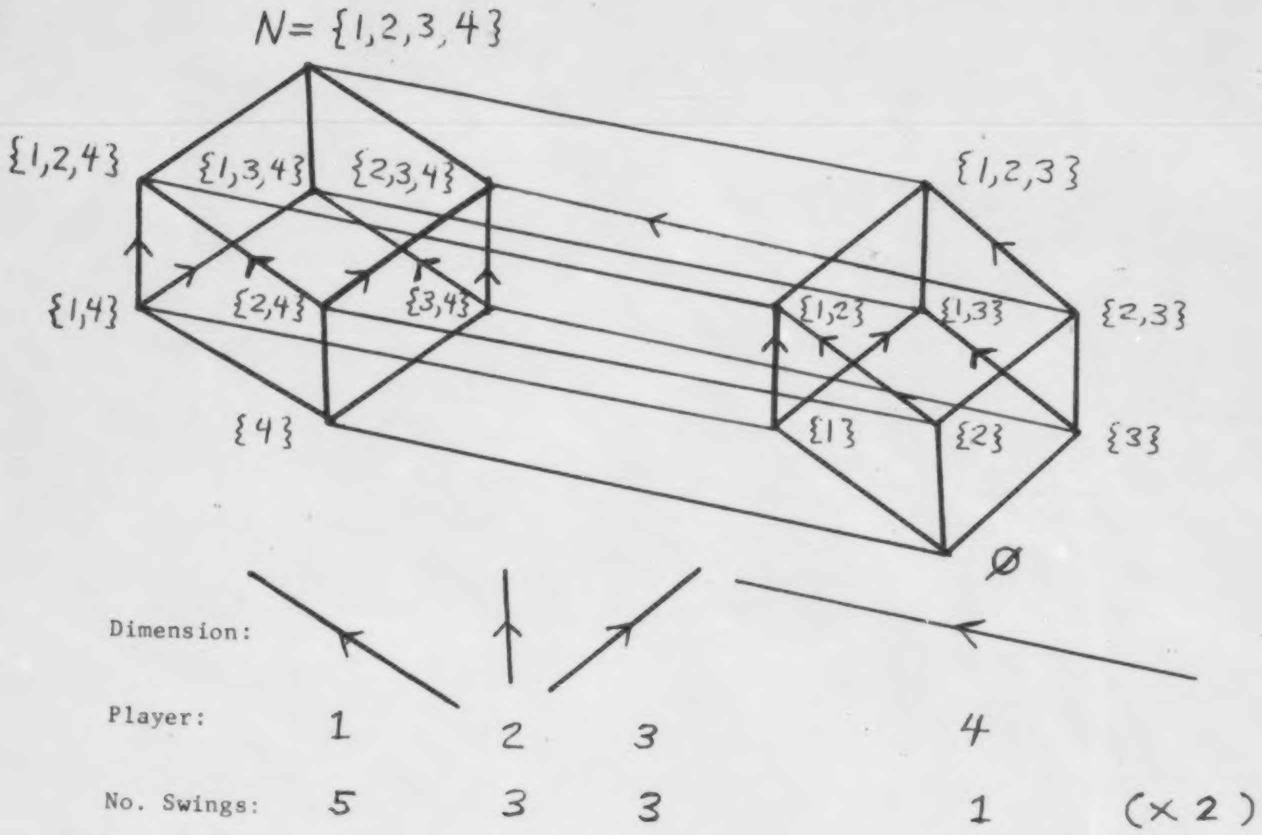

N= {1,2,3,4}
{1,2,4}
{1,3,4}
{2,3,4}
{1,2,3}
{1,4}
{2,4}
{3,4}
{1,2}
{1,3}
{2,3}
{4}
{1}
{2}
{3}
∅
Dimension:
Player:
1
2
3
4
No. Swings:
5
3
3
1
(× 2)

FIGURE 3.6.1

3.7 Other Value Concepts. Several other value concepts have been suggested for investigating power. The interested reader can consult the references listed on page 329 of Banzhaf (1965) and the discussion in Riker (1964), as well as the book by Bell et al. (1969). A different normalization of the Banzhaf index is used by Coleman (1971a), whereas Dubey and Shapley (1975) concentrate more on the actual probabilities $b_i/2^n$.

4. ADDITIONAL EXAMPLES

4.0. Introduction. Some examples and exercises for voting games which are usually somewhat more complicated than those given in Section 2 are presented in this Section, and some of these are taken from real-world situations.

4.1. The United Nations Security Council. The U. N. Security Council consists of the "big five" countries (France, Great Britian, the People's Republic of China, the Soviet Union, and the United States) who are permanent members, and each individually has veto power; plus ten "small" countries whose membership rotates. It takes nine votes, the "big five" plus at least four others, to carry an issue. This gives the game $[39;7,7,7,7,7,1,1,1,1,1,1,1,1,1,1]$. Note that a "small" country i can be pivotal in a winning coalition S if and only if S contains exactly nine countries including the five permanent members. There are $\frac{9!}{3!6!}$ such different S which contain i, and the corresponding coefficient in the Shapley-Shubik formula for this 15-person game is $\frac{(9-1)!(15-9)!}{15!}$. The product of these two numbers gives the power index $\varphi_s = .001863$ for any nonpermanent member. A member from the "big-five" has index $\varphi_b = (1-10\varphi_s)/5 = .1963$. It would be of interest to compute the Banzhaf index for this Council, and to compare these indices with those for the members of the old Security Council before 1963 which was $[27;5,5,5,5,5,1,1,1,1,1,1]$ (answer for φ: $6\varphi_s = 1/77$ and $5\varphi_b = 76/77$); or with a potential future Council which adds Japan and/or West Germany as a permanent member, also holding veto power. (The latter two countries now contribute more financially to the U. N. than do Britain or France.) A much more ambitious project is to compute φ or β for the present combined game in the Security Council and the General Assembly, as was done in Brams (1975) and Schwödiauer (1968). (See Section 4.3.)

4.2. Some Other Assemblies. Several other existing assemblies or committees can be found which are weighted voting systems, and it is left as an exercise to express these as weighted majority games and to determine the values for φ and β.

The Board of Estimates for New York City has had eight members with voting strength as indicated in Table 4.2.1.

NEW YORK CITY BOARD OF ESTIMATES

Member	No. of Votes Former System	Present System
Mayor	3	4
Controller	3	4
Council President	3	4
Brooklyn Borough President	2	2
Manhattan Borough President	2	2
Bronx Borough President	1	2
Richmond Borough President	1	2
Queens Borough President	1	2

TABLE 4.2.1

The Israeli Knesset in late 1965 had the division indicated in Table 4.2.2, taken from left to right on the political spectrum, as described by Owen (1971).

1965 ISRAELI KNESSET

Party	No. of Seats
New Communists	3
Communists	1
Poaelei Aguda	2
Agudat Israel	4
National Religious	11
Mapam	8
The Alignment (Mapai-Ahdut)	49
Haolan Haze	1
Independent Liberals	5
Rafi	10
Gahal	26

TABLE 4.2.2

The Tokyo Metropolitan Assembly had elections in 1973 with the results as indicated in Table 4.2.3.

TOKYO

Party	No. of Seats
Liberal-Democrats	51
Komeito	26
Communists	24
Socialists	20
Democratic-Socialists	2
Independents	2

TABLE 4.2.3

The 1973 general elections in Sweden ended in a virtual "tie" with a total of 175 seats between the ruling Social Democrats (157) and Communists (18), as well as for the three opposition parties: Center Party (90), Moderates (51) and Liberals (34).

The State Senate in Hawaii, as reported in Banzhaf (1966), had one district in the State with one legislator, one district with two, two districts with three each, and four districts with four legislators from each of them. He also studied the Arkansas House, the Georgia Senate, the Texas House, and the Wyoming Senate. Consider the power of the various districts under the assumption that all delegates from the same district vote the same way on any question. This multi-member district situation is studied in more depth in Section 5.3.

4.3. Some Large Games. For the 9-person game $[7;5,1,1,1,1,1,1,1,1]$ with one strong player, one can compute that $\varphi = (10,1,1,1,1,1,1,1,1)/18$. More generally, for the n-person game $[q;w_1,1,1,\dots,1]$, where w_1 is an integer such that $w_1 < q \leq n-1$, $\varphi = ((n-1)w_1, n-w_1, n-w_1, \dots, n-w_1)/n(n-1)$. Note that player 1 occupies each position in a random ordering of the n players with equal probability, and so his value is obtained by merely counting the number of times he is in pivotal position and dividing by n. Note that for small n, $\varphi_1 = w_1/n$ is somewhat larger than the first player's share of the weights which is $w_1/w = w_1/(w_1+n-1)$, where $w = \sum_{i \in N} w_i$. But for large n, φ_i approaches player i's share of the weights. Limiting results have been obtained for φ for several classes of games of this sort by Shapley and Shapiro (1960), Milnor and Shapley (1961), and Shapley (1961); and a recent book by Aumann and Shapley (1974) describes a whole value theory for various infinite games. Similar types of limiting results have also been obtained by Dubey and Shapley (1975) for the Banzhaf index.

In some decision procedures in Australia, each one of the six states has one vote, and the federal government has two votes plus the ability to break ties, when a simple majority is necessary to win. This can be expressed as a weighted majority game of the type just mentioned, and its φ and β can be determined. In the United Nations General Assembly with a required two-thirds majority one has the 138-person game $[94;3,1,1,\ldots,1]$ since the Soviet Union has three votes.

5. MINI PROJECTS

5.0. Introduction. This section includes a collection of relatively small projects, any one of which can be undertaken in a class, or by small subgroups of a class, over a period of two or three weeks. These illustrate applications of the power index concept and are suitable for a class which does not have the prerequisites, time or computer facility in order to pursue the theory or computation required in the next two sections. The calculations for these mini projects can be done by hand if necessary, and can thus be used as a warm-up or for numerical check against any computer programs developed in the following.

5.1. The Proposed Canadian Constitutional Amendment Scheme. A method for amending the Canadian constitution was proposed in 1971 which involved passage by the federal government as well as a ratifying procedure by the ten provinces. Our immediate objective is to investigate the voting powers exhibited in only this latter ten-person game between the provinces, and to compare the results with the provincial populations.

The winning coalitions or those with veto power can be described as follows. In order for passage, approval is required of

(a) any province which has (or ever had) 25% of the population,

(b) at least two of the four Atlantic provinces, and

(c) at least two of the four western provinces which currently contain together at least 50% of the total western population.

Using current population figures this means that veto power is held by

(d) Ontario (O) and Quebec (Q),

(e) any three of the four Atlantic (A) provinces (New Brunswick (NB), Nova Scotia (NS), Prince Edward Island (PEI), and Newfoundland (N)),

(f) British Columbia (BC) plus any one of the three prairie (P) provinces (Alberta (AL), Saskatchewan (S), and Manitoba (M)), and

(g) the three prairie provinces taken together.

The various types of winning coalitions are listed in Table 5.1.1. Ontario and Quebec are not listed since they are in every winning coalition, but are included in the number s of players in coalition S.

WINNING PROVINCIAL COALITIONS

Type	S	s	No. of such S
1	1P, 2A, BC	6	18
2	2P, 2A, BC	7	18
3	3P, 2A	7	6
4	1P, 3A, BC	7	12
5	3P, 2A, BC	8	6
6	2P, 3A, BC	8	12
7	3P, 3A	8	4
8	1P, 4A, BC	8	3
9	3P, 3A, BC	9	4
10	2P, 4A, BC	9	3
11	3P, 4A	9	1
12	3P, 4A, BC	10	1
		Total:	88

TABLE 5.1.1.

The numbers in the last column of Table 5.1.1 can be used with the formula for the Shapley value (Section 3.3) to compute φ for each of the provinces as well as regional groups, and these are shown in Table 5.1.2. For example,

$$\varphi_O = [18(5!4!) + 36(6!3!) + 25(7!2!) + 8(8!1!) + 1(9!0!)] \div (10!) = 53/168,$$

and

$$\varphi_M = [6(5!4!) + 10(6!3!) + 5(7!2!) + 1(8!1!)] \div (10!) = 1/24.$$

SHAPLEY-SHUBIK INDEX FOR PROVINCES

Province	φ(in %)	% Population	φ/Population
BC	12.50	9.38	1.334
AL	4.17	7.33	0.570
S	4.17	4.79	0.872
M	4.17	4.82	0.865
(4 Western)	(25.01)	(26.32)	(0.952)
O	31.55	34.85	0.905
Q	31.55	28.94	1.092
NB	2.98	3.09	0.965
NS	2.98	3.79	0.786
PEI	2.98	0.54	5.53
N	2.98	2.47	1.208
(3 Atlantic)	(11.92)	(9.89)	(1.206)

TABLE 5.1.2

The calculations for φ described above were made by D. R. Miller (1973), and he obtained his population figures for the ten-province area from the Canadian Almanac and Directory 1972 (Toronto, 1972). There are several extensions of his work which would also be of interest.

(i) Verify the results for φ which are given in Tables 5.1.1 and 5.1.2.

(ii) Determine the Banzhaf value β for this game between the provinces, and compare it to φ. Also compute the ration β(in %) ÷ % Population. These answers, as computed by Cornell University student Eleanor Walther in May of 1974 and by D. R. Miller (1974) and others are given in Table 5.1.3. Note that $\varphi_{AL} > \varphi_{NB}$ but that $\beta_{NB} > \beta_{AL}$, and thus this "simple" game is not a weighted majority game.

(iii) Find φ and β for a revised scheme which gives veto power to {BC,AL} and {BC,S,M} instead of (f) above, and thus more power to oil-rich Alberta. (Miller (1974).)

BANZHAF INDEX FOR PROVINCES

Province	β(in %)	β/Population
BC	16.34	1.74
AL	5.45	0.74
S	5.45	1.14
M	5.45	1.13
(4 Western)	(32.68)	(1.24)
O	21.78	0.62
Q	21.78	0.75
NB	5.94	1.92
NS	5.94	1.57
PEI	5.94	11.00
N	5.94	2.41
(4 Atlantic)	(23.76)	(2.40)

TABLE 5.1.3

(iv) Use the "square root" result in Sections 6.2 and 7.3 to determine the relative β for the individual voters (assuming all vote) in the voting game played within their own provinces. Then compute an approximate relative voting power of each citizen in the composed national game by multiplying these latter values by the corresponding φ or β for his province as given in (i) or (ii).

The mathematical results leave room for a good deal of discussion. Some have criticized this scheme (before seeing Miller's results) for giving BC too much power. Is the low figure for AL and the high one for PEI justifiable? What are some reasonable alternatives to the proposed scheme? One would wish to avoid making PEI a dummy, for example. A brief discussion along these lines is given in Miller's paper.

5.2. Power in a Nine-Man Body. A colleague related to me that he was a member of a nine-man committee on which he and four others formed a coalition. This bloc of five agreed covertly that they would all vote the same way on any issue before the full committee (except for a few deviations for diversionary reasons on minor issues or when others not in their coalition also supported their view), and thus a majority of three or more from this coalition effectively controlled the full committee. So each man in this minimal winning coalition increased his voting power from 1/9th to 1/5th. At the same time, however, my friend was wondering whether three of the other four in his coalition had not formed another secret three-man subcoalition without him, which in turn was controlling the whole committee. It would thus become possible to win on a ballot on which there were really only two favorable votes, and six members would be effectively dummies.

In cases like this, as well as others, it is of interest to investigate the power inherent in the cosets in various partitions of a nine-man body. One might also be concerned with changes in power caused by realigning such blocs, and whether the potential gains to counter-organization come from existing blocs or from the nonaligned individuals. The Shapley values for the various sized coalitions in the different types of partitions of a nine-man body have been determined by Krislov (1963) and listed in Table 5.2.1 for the case of a simple majority quota.

Some similar projects along these lines could be undertaken such as the following.

(i) Compute the Banzhaf value β for all subsets in all partitions of a nine-man body.

(ii) Determine φ and β for subsets of committees of sizes different than nine, e.g., of size 5, 7, 8, or 11; and denote the gains in new power ratios which accrue from coalition building.

(iii) Consider the previous cases for quotas other than a simple majority, e.g., three-fifths or two-thirds.

NINE-MAN COMMITTEE

Partition Type	φ
5,4 (etc.)	(1,0) (etc.)
4,4,1	(1,1,1)/3
4,3,2	(1,1,1)/3
4,3,1,1	(3,1,1,1)/6
4,2,2,1	(3,1,1,1)/6
4,2,1,1,1	(6,1,1,1,1)/10
4,1,1,1,1,1	(10,1,1,1,1,1)/15
3,3,3	(1,1,1)/3
3,3,2,1	(1,1,1,0)/3
3,3,1,1,1	(9,9,4,4,4)/30
3,2,2,2	(3,1,1,1)/6
3,2,2,1,1	(4,2,2,1,1)/10
3,2,1,1,1,1	(4,2,1,1,1,1)/10
3,1,1,1,1,1,1	(9,2,2,2,2,2,2)/21
2,2,2,2,1	(1,1,1,1,1)/5
2,2,2,1,1,1	(7,7,7,3,3,3)/30
2,2,1,1,1,1,1	(25,25,11,11,11,11,11)/105
2,1,1,1,1,1,1,1	(7,3,3,3,3,3,3,3)/28
1,1,1,1,1,1,1,1,1	(1,1,1,1,1,1,1,1,1)/9

TABLE 5.2.1

When one or more individuals defect from a coalition and migrate into another one (perhaps a new singleton), the realigned coalition structure (i.e., partition) will normally exhibit new power indices for the individuals. If "political man" attempts to maximize his voting power as "financial man" attempts to maximize his profits; then one might expect that such changes, e.g., individuals between political parties, might indicate an increase in the migrator's power or the total power of the party he joins, or that migrations are more common among those with little power to begin with. Riker (1959) studied these hypotheses in the case of the French National Assembly during the 1950's. During the period he considered there were many parties, some of which exhibited strict party discipline, as well as frequent migrations. For example, a typical Assembly was a majority game such as

$$[313;105,100,88,85,75,55,46,32,23,13,1,1,1].$$

Riker developed several formulas to test various power changes before and after migrations. Most of the results were rather ambiguous regarding his hypotheses. It appears that such changes were more for ideological reasons than attempts to gain power, but the situation may have been much too complicated for them to even sense, much less calculate, such indices.

Returning to our nine-man body, we see that several powerful coalitions are quite unstable in the sense that a defector can gain significantly. For example, a partition of type 5,4 seems highly unstable, since a change to type 4,4,1 reduces the power of the five-man coalition from 1 to 1/3 while increasing the swing-man's own share of the power from 1/5 to 1/3.

Some of the notions from Luce's theory of ψ-stability relate to models with changes in partitions. See Chapter 10 in R. D. Luce and H. Raiffa (1957).

Investigations into coalition stability and changes in power would be of interest. Brams and Affuso (1975) discuss power changes due to additional voters.
(iv) Use Table 5.2.1, or the results from (i) above, and Riker's paper as an outline to study stability and changes in power to the various individuals and groups in a nine-man committee in which migration is allowed. Similar studies can be done for committees of other sizes, using for example the results in (ii).
(v) Investigate such changes in power for some real-world body, e.g., some local college committee, or the U. S. Supreme Court where death and new appointments or other reasons bring about changes in the coalition structures as in the case of Justice Rutledge. The reference by Schubert (1959) is of some interest here.

5.3. Multi-Member Districts. In many voting assemblies, such as some state legislatures, the Electoral College, a university senate, etc., the voting representatives often come from districts which differ greatly in the number of citizens they have. It is often thought that any inequity created by differences in population can be balanced by electing different numbers of representatives from the various districts. For example, if one district in a state contains a city and has 4,000,000 people whereas all other districts have

1,000,000 inhabitants, then it is viewed as "fair" if the first district elects four times as many representatives at-large from its region as does each other district in the state. This seems to assume that the ability of an individual citizen to affect the election of his representatives varies inversely with population. It is shown in the following examples that this is not a valid assumption in terms of the Banzhaf power index. In fact, a citizen-voter's decreasing influence in electing his representatives varies inversely with the square root of the increasing population, and this will be demonstrated in general in Section 6.2.

Banzhaf (1966) has illustrated how one's influence varies with population for districts with only a few citizens and in the case when the elections are between just two opposing parties, and his example is shown in Table 5.3.1.

SOME MINIATURE DISTRICTS

District Symbol	No. of Voters	No. of Combin.	No. Combin. Marginal	% Combin. Marginal	Individ. % of Pop.		
	n	2^n	b	$100b/2^n$	$100/n$	$100/\sqrt{n}$	$nb/2^n$
A	3	8	4	50.0	33.3	57.7	1.50
B	5	32	12	37.5	20.0	44.7	1.88
C	7	128	40	31.2	14.3	37.7	2.18
D	9	512	140	27.4	11.1	33.3	2.46

TABLE 5.3.1

Note there that $b = 2\times(2m)!/(m!)^2$ where $m = (n-1)/2$. For example, when $n = 9$, the per cent of voting combinations in which a citizen is marginal (27.4) is closer to $100/\sqrt{n} = 33.3$ than it is to $100/n = 11.1$. For large n the reciprocal square root approximation, when properly normalized, is extremely accurate (see Section 6.2). So that one could argue in the example mentioned above, that the city with 4,000,000 inhabitants should only have twice the number of representatives as a district with 1,000,000 constituents. This type of argument can be made more forcefully when one considers the influence the representatives from large districts have in turn in the state legislature if they vote as an irrevocable bloc. The effect of so "composing" two such voting games is shown in a simple example below, and for the more realistic Electoral College in Section 7.

Consider a small "state" which is divided into seven districts in which the first district has a voting population of 9 and each of the other six districts has a voting population of 3. Assume that this large district elects 3 representatives at-large by choosing a group of three people from the same party who then vote the same way on any issue before the state legislature. Each one of the small districts elects one person to the legislature. Furthermore, assume that each representative acts as a true delegate and votes on each question the way in which the majority of his constituents would vote. We see from Table 5.3.1 that a citizen-voter from a small district can influence his or her representative's vote in 50% of the voting combinations, whereas the individual in the large district can influence the vote by the party of three legislators in only 27.4% of the combinations in his district's polls. However, in the legislature itself one has the weighted majority game $[5;3,1,1,1,1,1,1]$. One can show that the Banzhaf power index for this game is $\beta = (5,1,1,1,1,1,1)/11$. This is close to the Shapley-Shubik index of $\varphi = (9,2,2,2,2,2,2)/21$ which was computed in Section 4 for all such games with one large player. So a citizen-voter from the large district can influence his group of three representatives 27.4% (= 140/512) of the time, and they in turn can sway the legislature 78.1% (= 100/128) of the time, for a product of 21.4%. On the other hand, an individual from a small district influences his lone legislator 50% (= 4/8) of the time, and this delegate in turn changes the full legislature 15.6% (= 20/128) of the time, for a product of 7.8%. These products can approximate an individual's power in the "compound" game.

Previous examples had indicated that in a weighted voting situation the "big" players with the most votes often, but not always, have somewhat larger indices than their fractions of the total vote would indicate. In this section one observes further that if several delegates are elected at-large from various districts, then the influence of the individual voter in the large districts may be greatly exaggerated relative to the other regions when the number of delegates per district is proportional to its population. Banzhaf (1966) actually applied this sort of an analysis to five state legislatures. It would make an interesting project to investigate some additional examples which have such multi-member representation, under the assumption of bloc voting behavior.

(i) Investigate some local or state governmental body, board, agency or convention of citizens which has a multi-member representative structure. For

example, some school boards have a representation from various constituencies. The school boards in New York City would make an interesting study.
(ii) A college or university has several schools, divisions or departments of various sizes. Investigate how to construct a representative council for some such structure, for example, a mathematics departmental or divisional executive committee which has "fair" representation from each interest or specialty group.

5.4. Some Current Parliaments. A review of recent newspapers and newsmagazines will give interesting data concerning current elections and national assemblies. Especially in the year 1974, when there are several minority governments in the Western World in which a winning coalition of parties must form in order to obtain a ruling government. Some countries also have five to ten different parties in their parliaments, and this requires a reasonable sized computational exercise. Data for four countries are listed below; and one can compute φ and β for these assemblies under the assumption of strict party-line votes. This latter assumption has some validity in at least the formative stages of a government, and even at later stages in those countries in which strict party discipline exists.

The following data are from the spring of 1974; and even some of it should be checked further or extended in order to be completely reliable. Such figures are frequently in flux, and one is dealing here with rather dynamical situations. For example, some members of the British Parliament soon lost their votes because they had taken leadership roles, etc. The "other" category could also be broken into three singletons. The newspaper reports on Israel were inconsistent and difficult to sort out, and so the following data should be validated. The figures on Denmark do not include a small number of representatives from Greenland and the Faroe Islands. If one wished to do a current and accurate investigation, he should first verify and refine the following data. It is also interesting how most election results in the newspapers are reported as the per cent of the popular vote taken by each party (or often just the major parties) rather than the number of seats won; whereas the latter results are the more critical ones, and they are not always easily obtainable from the former per cents.

(i) Luxembourg:

	No. of Seats	
Party	Early 1974	Mid 1974
Christian Democrats	21	18
Socialist	12	17
Splinter Socialists	6	5
Liberals	11	14
Communists	6	5
w	56	59
q	29	30

TABLE 5.4(i)

(ii) Great Britian:

Party	No. of Seats
Labor	301
Conservatives	296
Liberals	14
Irish Unionists	11
Scottish Nationalists	7
Others	3
Welsh Plaid Cymru	2
Irish Catholics	1
w	635

TABLE 5.4(ii)

(iii) Israel:

Party	No. of Seats
Labor	51
Likud	39
National Religious	10
Independent Liberals	4
New Communists	4
Orthodox Agudat Israel	3
Agudat Israel Workers	3
Arab Affiliates	3
Civil Rights	2
Moked	1
w	120

TABLE 5.4(iii)

(iv) Denmark:

Party	No. of Seats	Ordinate*
Communists	6	.00
Socialist Peoples	11	.18
Social Democrats	46	.37
Center Democrats	14	.50
Justice League	5	.54
Radical Left	20	.58
Left (Liberals)	22	.64
Christian Peoples	7	.79
Conservative Peoples	16	.84
Progress	28	1.00
w	175	

TABLE 5.4(iv)

Even in countries which have a one-party majority government, one may be able to study the power in various interest groups within this party or the assembly as a whole. See, for example, the study by Lieserson (1968) on Japan.

5.5. Cornell University Senate. A broadly representative University Senate has been in operation at Cornell in recent years, and it has significant powers theoretically over events and budgets in important areas such as student life and activities. The size of this Senate as well as of each of its six major groups has varied, and these numbers are given in Table 5.5.1. The number in the Employee's category was increased by eight between 1971 and 1972 for several reasons. Some of these were perhaps more idealistic, such as support for labor in general or because of the large number of members in this class. (At Cornell there are nearly four people in the mostly invisible "supply line" for each professor at the "front.") Other reasons were more practical, such as the fact that the employees were critically involved in the consequences of Senate decisions, and usually it was desirable to have at least one such overworked employee senator on most of the two dozen or so committees and subcommittees. The total size of the Senate was reduced by 1974 in attempts to improve efficiency and because of the large absentee

---

* These ordinates are an estimate of the parties positions on the political spectrum as measured from left to right, and they will be referred to in Section 8.1.

CORNELL SENATE

Constituency	1971	φ	β	1972	1974	No.
Students	60	9	4	60	40	15,500
Faculty	60	9	4	60	40	1,550
Employees	5	6	3	13	10	5,000
Alumni	2	2	1	2	1	130,000
Administration	2	2	1	2	2	9
Non-Prof. Acad.	3	2	1	3	2	550
w	132	(÷)	(÷)	140	95	
q	67	(30)	(14)	71	48	

TABLE 5.5.1

problems. The Shapley-Shubik and Banzhaf values for the 1971 Senate are shown in the Table. The correction for the benefit of the Employee group in the two later Senates may have been "overdone" however, since the result was three groups of equal power and three dummies. (Of course, some students knew right along that the Alumni, Administration (e.g., the Provost and vice presidents) and Non-Professional Academics (e.g., librarians, etc.) were "dummies.")

The assumption of bloc voting or "party-line" discipline within the six groups listed in the Table is clearly invalid in practice for this particular Senate, as one would expect in most college situations. Extremes of opinion within some of the large groups could be found on almost any issue. On the other hand, there may be specific local issues on which a partitioning of the Senate into various interest groups might make sense. For example, some groups may have strong feelings in favor or against athletics, whereas different groups may have intense "pro" or "anti" attitudes on Greek living units. Consistent voting patterns might also be discerned on topics such as women or minority students and affirmative action programs. Such data may lead to power analyses and prediction of outcomes on such individual legislation.

There are many interesting projects which one can undertake concerning university, college, school or departmental senates, and a few of these are suggested below. One could use the data in Table 5.5.1, or collect similar figures for his own institution. The numbers for the six constituencies shown in the right hand column in this Table, are for the 1971 Cornell Senate. Except for constitutional amendments and a few procedural matters, a simple majority is used for the quota.

(i) Verify the values of φ and β given in Table 5.5.1.

(ii) Do a "two-level" analysis using the φ and β for Table 5.5.1 (or your own college data) plus the "square-root" result in Sections 6.2 and 7.3 to estimate the power of an individual person (not senator) in the six groups.

(iii) If the committee structure of a senate is known, then one can investigate the power of an individual representative in the sequential process of getting a bill through committee and the senate. For example, as was the case in the first Cornell Senate, an alumnus on the powerful Executive Committee would increase his power as well as that of the Alumni group.

(iv) Given the number of student representatives a senate will have, determine a "fair" allocation of this number among the various colleges in a university, or departments in a college.

(v) One can design a senate from scratch by deciding what power the different individuals and/or groups should have and then setting the number of representatives from the various constituencies so as to best achieve this. (See the following Section for additional ideas on (iv) and (v).

(vi) The power of the various committees and/or interest blocs for a university board of trustees or state regents may be of interest, especially in the case of a rather large board as is the case for Cornell University.

5.6. Designing Representative Bodies. In most of the weighted voting problems discussed in this paper, one is given a well structured problem (i.e., the weights) and is asked to determine the corresponding voting power as measured by the Shapley-Shubik and/or Banzhaf indices. On the other hand, one is often faced with the opposite or reverse problem in practice; i.e., if one is designing a legislature, he is given the size of the districts and he must then determine the weights or number of representatives for each in order to achieve a certain given or desirable measure of power. Projects (iv) and (v) in the previous Section are of this type. Most work in this reverse direction uses less elegant or straightforward techniques. One often uses more ad hoc, self correcting, or guess-and-verify methods. Some approximation algorithms do exist, although they may not always be readily available (without a fee). There are also several substantial and unsolved mathematical problems along these lines, in addition to the practical ones; but the author will not go into further detail here on these more theoretical directions. One should remember, however, that there may be alternate types of solutions to his voting problem, e.g., reapportionment, restricting, or multi-member districts, in addition to finding the "correct" (unequal) weights for the lone representative from each district.

Let us return to the voting rules in the Nassau County Board of Supervisors which is described in Table 2.3.1 in Section 2.3. Recall that the 1958 and 1964 systems each had three dummy municipalities. In 1971 a new scheme was passed which merely changed the quota from a simple majority of 58 to 63 votes in order to pass a bill. (Actually, the County Charter in 1938 allows one vote per 10,000 residents, and thus Hempstead should currently have a weight of 72. However, since no town or city is allowed by this Charter to have more than one-half of the total vote of the board (i.e., $125 \div 2$), one subtracts ten votes from Hempstead to get 62 and takes w as 115). There are no dummies in the revised scheme, and the new values of φ and β, as computed by a Cornell student, Shigeo Muto, with $q = 63$, are shown in Table 5.6.1.

NASSAU COUNTY REVISITED

Municipality	No. of Votes	$60\ \varphi_i$	$54\ \beta_i$
Hempstead No. 1	31	17	15
Hempstead No. 2	31	17	15
North Hempstead	21	7	7
Oyster Bay	28	13	11
Glen Cove	2	3	3
Long Beach	2	3	3

TABLE 5.6.1

Note that the new power indices are not too extreme in comparison with the proportions of the population. Attempts to invalidate this latter scheme via the courts have failed. Related stories appeared in the New York Times on May 4, 9, and 12, 1971 and February 2, 1974.

(i) Verify the values of φ and β given in Table 5.6.1.

(ii) Another problem in the Nassau County scheme is that the two representatives from Hempstead are elected at-large from the full district, which gives rise to questions raised in Sections 6.2 and 7.4. Reconsider this scheme in light of this additional problem. It would also be of interest to study the present situation in Nassau County, since the above system was overturned by a referendum (New York Times, November 17, 1974).

In 1971 a suit was brought against Cortland County, New York which claimed that its one town-one vote voting scheme was unfair, and the state courts agreed that it was unconstitutional. A proportionately-weighted interim scheme was then set up, while the County undertood to devise a new valid apportionment plan. The new system, involving some redistricting, used the 1970 Decennial Census figures and the Banzhaf index for weighted voting. It has since been accepted by the courts, and took effect on January 1, 1974. The new law for the Cortland County Legislature gives one legislature to each of the 19 districts, and weights his vote as indicated in Table 5.6.2. There are different weights necessary for each of three cases, depending upon whether a simple majority, a three-fifths majority, or a two-thirds majority is required to pass an issue. The computation of the desired weights was done by Lee Papayanopoulos, a New York City based consultant, who has a computer program for obtaining approximate weights to satisfy equity in the sense of the Banzhaf index. (The County paid $2600 for his report, and he later collected $350 per day for his services in appearing in court.) A study of the Table can point out some "discrepancies" which should give rise to further questions. For example, the nine largest districts control 235 votes for more than a simple majority of 231, and yet they have only 49.2% of the population. Note also, for example, that in the two-thirds plan, Districts 15, 16, and 6 have one less weight than the slightly smaller Districts 5 and 19. Incidentally, there was some controversy in designing the current scheme about how to count the students at Cortland State.

(iii) Consider a corporation in which four stockholders have 60, 20, 10, and 10 per cent of the stock respectively, and for which the company charter calls for all measures to be passed by a simple majority vote of the stock. The directors, knowing that the majority stockholder is a dictator (in our mathematical sense), agree to modify the rules to have "voting power" (say as measured by φ and/or β) approximate the relative stock holdings. Two proposals (or is it three or four?) are put forward:

A. Majority rules except when all minority stockholders unify against the majority stockholder. (One may question whether this verbal statement means that the minority have merely veto power or are a winning coalition.)

B. Majority rules except when at least two-thirds of the minority stockholders' votes are opposed to the majority stockholder.

CORTLAND COUNTY LEGISLATURE

District Number	Population	% of Pop.	Weights when Majority Quota is 1/2	3/5	2/3	Case q = 1/2 % of Vote	% of Power
3	2716	5.918	31	40	38	6.739	5.859
7	2560	5.578	27	35	36	5.870	5.612
13	2517	5.484	26	34	34	5.652	5.482
14	2500	5.447	26	33	34	5.652	5.482
2	2478	5.399	25	33	33	5.435	5.358
8	2467	5.375	25	32	33	5.435	5.358
18	2460	5.360	25	33	33	5.435	5.358
15	2444	5.325	25	32	32	5.435	5.358
16	2442	5.321	25	32	32	5.435	5.358
6	2442	5.321	25	32	32	5.435	5.358
5	2440	5.317	25	32	33	5.435	5.358
19	2434	5.304	24	32	33	5.217	5.238
12	2442	5.277	24	32	32	5.217	5.238
1	2406	5.243	24	31	31	5.217	5.238
17	2402	5.234	24	31	31	5.217	5.238
4	2284	4.977	22	28	29	4.783	4.916
9	2187	4.765	20	26	25	4.348	4.735
10	2153	4.691	19	25	24	4.130	4.727
11	2140	4.663	18	25	23	3.913	4.727
Totals	45894	100.000	460	598	598	100.000	100.000

TABLE 5.6.2

(iv) One might wish to reconsider the projects (iv) and (v) in the previous Section 5.6 in light of the discussion in the present Section.

(v) It would be of interest to undertake a project to redesign some local governmental body. For example, there are still several counties in New York State with unconstitutional systems which are in need of revision. This redesigning has been done for several other counties in New York, and an excellent report on this through 1968 is given in Johnson (1969), and a more recent study is presented by Imrie (1973).

5.7. The U. S. Congress. Several investigations have been made into measures of power for various sorts of legislatures. In a unicameral body in which a winning coalition depends only upon the number of members it contains, each legislator theoretically has equal power. More interesting problems occur in multicameral bodies or "composed" games, in which various majorities in several hourses of different size may be required for passage. An interesting case is the U.S. Congress where a bill requires a majority in the House and the Senate plus the President's signature, or a two-thirds majority of both houses in the case of a veto override. Shapley and Shubik (1954) have approximated φ as: 5/12 for the House of Representatives as a whole, 5/12 for the Senate, and 1/6 for the President. Several other studies have since been done of the U. S. Congress, and one can refer to Section 12.1 of Luce and Raiffa (1957), to Chapters 14 (especially page 215) and 16 (pages 245-250) of Rapoport (1970), as well as to the references given there for reports on the early work done in this area. Also of interest is Chapter 11 by Luce and Rogow in Shubik (1964), and the paper by Riker and Niemi (1962). Several types of similar studies should be of interest.

(i) Investigate the various interest blocs in some legislature like the U. S. Congress, at least for the case of some particular interest area, and then determine the voting power of the various blocs.

(ii) Do a power analysis for some state legislative system.

(iii) Consider the "committee system" which exists for some legislative body such as the U. S. Congress or a University Senate, and determine power indices under the assumption that a bill must pass in the committee before it reaches the floor. See the paper by Brams and Papayanopoulos (1974) on the Congress.

5.8. Some Mathematical Problems. By now it should be clear to the mathematical reader that there are a great number of mathematical questions and relations which can be pursued, as class projects or basic research, concerning our weighted voting games and power indices. Although our main goal has been to present interesting or relevant applications suitable for mathematical modeling, a few more theoretical sorts of problems are mentioned very briefly in this Section. Although there are a host of such interesting questions, only a couple are suggested. No current survey of the state of this field is intended, and no detailed references are given. Although the suggested problems are often very difficult for general n-person voting situations, suitable class projects result if they are restricted to the cases of $n = 3$ or 4.

For an arbitrary n and given q, the set of all possible weighted majority games $[q;w_1,w_2,\ldots,w_n]$ can be considered as the points in an $(n-1)$-dimensional simplex W embedded in n-space. One can normalize the weights so that they sum to $w = 1$. To each game there corresponds a particular collection of winning (or minimal winning) coalitions in the lattice of all subsets 2^N of the voter set $N = \{1,2,\ldots,n\}$. Each such collection in turn gives rise to a Shapley-Shubik index φ, as well as a Banzhaf index β. Although the set W is infinite, the resulting number of collections of winning coalitions as well as the collections Φ and B of realizable indices φ and β are finite. Since any φ_i can be written as an integer divided by $n!$, and any β_i can, for each particular weight vector, be written as an integer over the total number of marginal situations for all n players, one can investigate how W partitions into equivalent games, i.e., games which give the same φ (or β). One can pick a particular q, or more generally, partition $W \times [0,1]$ as q varies over the interval 0 to 1 (or just over $(1/2,1]$ if one wishes to maintain our earlier assumption). One can see that not all fractions with a denominator of $n!$ can be realized as a φ_i.

It is natural to inquire as to which games have $\varphi = \beta$, and when is φ or β proportional to the weight vector $(w_1,w_2,\ldots,w_n)$. A few examples in the case of φ, are given on pages 208-209 of Riker and Shapley (1968).

If an index φ, e.g., $\varphi = (3,1,1,1)/6$, is considered as a weight vector of a new game, then this game will not in turn have the same index φ. Furthermore, not all voting situations (in the sense of minimal winning coalitions) can be represented as weighted majority games. One can consider this question in the case of the Canadian Constitutional Amendment Scheme (Section 5.1); or for the seven-point plane projective geometry discussed on pages 469-470 in von Neumann and Morgenstern (1953), i.e., where the minimal winning coalitions consist of the three points (or voters) which lie on a particular one of the seven lines. One interested in mathematical questions about weights should consult Isbell (1964) and Lapidot (1972) and the references they have.

Various methods for composing or combining games can be introduced. The corresponding power indices do not normally compose in any straightforward manner. For example, consider the situation [3;2,1,1] with $\varphi = (4,1,1)/6$, in which the third player is really a group of three players whose vote in turn is determined by majority rule. So this "player" really plays the game [2;1,1,1] with $\varphi = (1,1,1)/3$ before he votes in the larger situation. We can view this succession of votes as the one game [6;4,2,1,1,1] which has $\varphi = (39,9,4,4,4)/60$ and $\beta = (6,2,1,1,1)/11$. So, for example φ_1 has changed from 2/3 in the original situation to 39/60 in the composite game.

Some results on the average value of φ_i in a given game when the quota q is allowed to vary are discussed in Mann and Shapley (1960), e.g., on page 16.

6. COMPUTATIONAL AIDS

6.0. Introduction. In order to obtain the power indices for games with many players and irregular weights, such as the Electoral College discussed in Section 7, it is rather essential to have some computational devices, as well as machine facilities and programs to implement them. Mann and Shapley (1960) used some Monte Carlo techniques to obtain φ for the Electoral College, but greater accuracy was desired. Section 6.1 describes some mathematical tricks, suggested by Mann and Shapley (1962), and by David G. Cantor, which simplify the work of calculation; and an actual algorithm for performing this task is referred to. Section 6.2 describes an approximation method using Stirling's formula for computing the number of voting combinations in which an individual voter from a large district is marginal. This technique can be used to compare the relative influences in the sense of Banzhaf of two citizens from different sized districts with equal representation as discussed in Section 5.3, as well as for two individuals from different states in the U. S. Electoral College as is done in Section 7. Section 6.3 mentions an extension of the Shapley value due to G. Owen which gives a good approximation technique for determining φ and β for some large voting games.

6.1. Computing the Shapley-Shubik Index. The formula developed above for φ was expressed as

$$\varphi_i = \sum \frac{(s-1)!(n-s)!}{n!}$$

where the summation is taken over all coalitions S of s players in which i is pivotal. If one lets

c^i_{js} = Number of ways in which s players, other than i, can have a sum of weights equal to j,

then the formula, in the case of the n-person game $[q;w_1,w_2,\ldots,w_n]$, can be written as

$$\varphi_i = \frac{1}{n} \sum_{s=0}^{n-1} \binom{n-1}{s}^{-1} \sum_{j=q-w_i}^{q-1} c^i_{js},$$

where the numbers

$$\binom{n-1}{s}^{-1} = \frac{s!(n-1-s)!}{(n-1)!}$$

are the binomial coefficients, and can be read off the (n-1)st row of the Pascal triangle. This latter coefficient, when divided by n, is the number of ways, in which the s players in a given coalition S which join prior to i and the remaining n-1-s players in $N-\{i\}-S$ which follow i, can be ordered. If, and only if, the weights of the s members in S sum to a j in the range $q - w_i \leq j \leq q - 1$ does player i have enough weight w_i in order to join S and thereby first reach or exceed the quota q. The c^i_{js} then give the number of such distinct coalitions for each j and s. The number of terms in this double summation for φ_i is not excessive for machine calculations, and thus the major task is to determine the values for the c^i_{js}.

As related by Mann and Shapley (1962), Cantor's contribution was to point our that the c^i_{js} can be readily obtained from the generating function

$$f_i(x,y) = \Pi(1+x^{w_k}y)$$

where the product is taken over all $k \epsilon N-\{i\}$, since the desired c^i_{js} are merely the coefficients of the corresponding x^jy^s terms. The proof of this is left as an exercise. For any i, $f_i(x,y)$ can be obtained from the full n-fold product $\Pi(1+x^{w_k}y)$ divided by $(1+x^{w_i}y)$.

The c^i_{js} can also be found as the elements of a matrix C^i of integers, and for each player i this matrix can be generated inductively as follows. Define $C^{(0)}$ so that $c^{(0)}_{00} = 1$ and all other $c^{(0)}_{js} = 0$. Then $C^{(r)}$ is obtained from $C^{(r-1)}$ by the relation

$$c_{js}^{(r)} = c_{js}^{(r-1)} + c_{j-w_p,s-1}^{(r-1)}$$

where the last term is taken as 0 when either of its subscripts is negative, and the successive w_p stand for the weights of the distinct players in $N-\{i\}$. After $n-1$ iterations one gets $C^{(n-1)} = C^i$. One can obtain $C^{(n)}$ by taking all $r \in N$, and then get each C^i by subtracting once, i.e., by reversing our recursive relation above. Also see Brams and Affuso (1975).

Consider the example $[q;w_1,w_2,w_3,w_4] = [6;4,3,2,1]$, and compute φ_1. Only the nonzero elements of our matrices will be given. Beginning with $C^{(0)}$ with $c_{00}^{(0)} = 1$, and using $c_{js}^{(1)} = c_{js}^{(0)} + c_{j-w_2,s-1}^{(0)}$, gives $C^{(1)}$ with $c_{00}^{(1)} = c_{31}^{(1)} = 1$. Then $c_{js}^{(2)} = c_{js}^{(1)} + c_{j-w_3,s-1}^{(1)}$ gives $C^{(2)}$ with $c_{00}^{(2)} = c_{21}^{(2)} = c_{31}^{(2)} = c_{52}^{(2)} = 1$. And $c_{js}^{(3)} = c_{js}^{(2)} + c_{j-w_4,s-1}^{(2)}$ gives $C^{(3)}$ with $c_{00}^{(3)} = c_{11}^{(3)} = c_{21}^{(3)} = c_{31}^{(3)} = c_{32}^{(3)} = c_{42}^{(3)} = c_{52}^{(3)} = c_{63}^{(3)} = 1$. Our formula restricts j to the range $2 \leq j \leq 5$, and gives

$$\varphi_1 = \frac{1!2!}{4!} \cdot 2 + \frac{2!1!}{4!} \cdot 3 = \frac{5}{12} .$$

Not all of the above mentioned matrices need be stored in the computer, and only the ranges $0 \leq j \leq q - 1$ and $0 \leq s \leq (n-1)/2$ are necessary since $C^{(n-1)}$ is "symmetric" in the sense that $c_{js} = c_{w-w_p-j,n-1-s}$.

A detailed algorithm suitable for computer programming and designed to calculate φ using the approach described above has recently been written by Boyce and Cross (1973). These $c_{js}^{(r)}$ can also prove useful in computing β.

6.2. The Number of Marginal Combinations. Since the number of voting combinations, 2^n, and the number of marginal combinations, involving factors of the order $n!$, are much too large for even machine calculation, it is necessary to develop simpler approximating devices in order to handle realistic sized voting districts. Assume that a district has n voters where n is taken as odd, and that an election is won by a simple majority, that is, by $(n+1)/2$ votes. A particular voter is marginal whenever the other $n-1$ players divide into exactly one half for an issue and the other half against it. The number of such equal divisions or combinations is given by

$$\frac{2(n-1)!}{(\frac{n-1}{2})!(\frac{n-1}{2})!}$$

where the 2 in the numerator comes from assigning either side in a division the "yes" vote. One can obtain the relative influence of each player in the sense of Banzhaf by dividing this fraction by 2^n. Using Stirling's formula (which appears in many advanced calculus texts),

$$m! \approx \sqrt{2\pi m}\ m^m e^{-m}$$

one can reduce this new fraction to

$$\frac{2(n-1)!}{[(\frac{n-1}{2})!]^2} \times \frac{1}{2^n} \approx \frac{}{\sqrt{2\pi(n-1)}} .$$

If we assume that all voting combinations are equally likely, then this is the fraction of elections in which a particular voter i is decisive, and it varies inversely with the square root of the population. Note that since each voter has equal power, one must get $\beta_i = 1/n$; but β_i does not give the fraction of elections in which player i is in the critical position.

This result now justifies some of the assertions made in Section 5.3, and will prove useful in Section 7. It shows that the discrimination or the deprivation built into voting districts relates to this square root factor, if one accepts the Banzhaf analysis. This approximation is a good one for large districts since Stirling's formula is accurate, for example, to one part in one hundred for $m = 100$. Note that this approximation given by Stirling's formula is in the asymptotic sense, i.e., the ratio of the two sides of the equation approaches one for large m; the actual gap between $m!$ and the right hand side actually increases with m. Another discussion of this "square root effect" is given in pages 83-85 of Shapley (1973).

In a few footnotes, Banzhaf (1965, page 335; 1968, page 312) mentions some people who assisted him in calculating β for some of his applications. One computer program by M. Sackson was copyrighted, and leads one to wonder if mathematicians can increase their incomes by copyrighting their important theorems. In Section 5.6, reference was made to Lee Papayanopoulos who assisted Cortland County, New York in designing a weighted voting system using the Banzhaf approach. Johnson (1969) also described some of his computer programs for determining proper weights.

Another view of marginality is given in Brams and Davis (1973, 1974).

6.3. A Computation Method for Large Games. Owen (1972) introduced a generalization of the Shapley value in the context of general n-person games, and a specific application of his extension is an approximation technique for determining the Shapley-Shubik index. His approach is presented on pages P72-P73 in his paper, and it is quite accurate for games with a large number n of voters if each of the weights w_i is relatively small. It need not be too precise when there is exactly one relatively large weight. Owen's method makes use of the normal probability distribution and numerical integration. However, the actual computations are much easier than the exact values as obtained via Section 6.1 in the case of large games such as the Electoral College given in Section 7; and Owen (1975a) has made such calculations for the situation in 1972. Owen's approach is also discussed in pages 78-83 of Shapley (1973), along with an application to a proposed New Mexico weighted voting scheme. Owen (1975b) has also shown how his approach can be used to approximate β.

7. THE ELECTORAL COLLEGE

7.0. Introduction. Previous examples have shown that weighted voting systems occur rather frequently in political situations. This Section will be concerned with the familiar process of electing a president of the United States. This is a three level process in which the citizens in each state vote to determine their choice (the state game), who then receives all of that state's weighted vote in the Electoral College (the national game). If the latter vote is indecisive, then the final choice is made in Congress (the Congressional game). This procedure will be described in more detail in Section 7.1; and the national game, state games, and the combination of these games will be analyzed with respect to voting power in Sections 7.2, 7.3, and 7.4, respectively. Some criticisms of this election process are mentioned in Section 7.5, and some possible alternatives to it are also described there.

The American Electoral College provides a well-known real-world institution which has many nonobvious quantitative aspects despite its apparently transparent structure. It is a topic of at least periodic interest, and major changes in it have frequently been proposed. The mathematical complexity of this election process, and of the proposed alternatives to it, make it appropriate for a classroom project for students with some computing capabilities available.

7.1. Electing a President. Consider the indirect procedure by which an American president is elected. It begins with a national popular vote by all eligible citizens who choose to exercise their right. Each voter picks a particular candidate representing a certain political party. (We are not considering here the preliminary aspects concerning how the candidates survived state primary elections, party conventions, etc.) The candidate who wins the most votes in a given state, no matter how small his margin, will then receive all of that states weighted vote in the Electoral College. This is referred to as the "winner take all" or "unit-vote" rule, which is imposed by state laws. The individual electors who actually cast the votes in the Electoral College are usually mere functionaries. This purely mechanical role is contrary to their view in the Constitution and has had a few defectors in recent years, and the legality of some such state laws may be questionable. In short, the popular vote for the president only determines a group of electors for each state who in turn are pledged to the winning candidate in their state.

The number of electors which each state has in the Electoral College is equal to its number of representatives in both houses of the Congress. Each state has two senators plus one to forty-five representatives, where this latter number depends upon its population as determined by the regular decennial census and resulting apportionment (which gives rise to another interesting mathematical problem which will be mentioned again in Section 8). Since there are 100 senators, 435 representatives, plus 3 electors from the District of Columbia, there are a total of 538 seats in the Electoral College. Any candidate who reaches or exceeds the simple majority of 270 votes is declared the new president.

In the event that no one receives a majority in the Electoral College, then the process goes to a third level and the deadlock is resolved in the Congress. This occurred in the elections of 1800 and 1824. The House of Representatives chooses the president from among the top three candidates in the Electoral College. (The Senate picks the vice-president from the top two candidates.) In the House, each state has only a single vote, and a majority of 26 is necessary to win. Note that the small states could have extraordinary power relative to their total population in this Congressional determination. The following mathematical analysis will not be concerned with this third level, or potential "tie-breaking" procedure, in the presidential elections.

At this third level some truly bizarre possibilities can occur. For example, at one time before the 1968 election it looked as though the order of finish in the Electoral College could be Nixon-Agnew, Wallace-LeMay, and Humphrey-Muskie with the leading Nixon ticket having less than a majority. This could have resulted in H. H. Humphrey as President and General Curtis LeMay as Vice President. In a fictitious novel, Russell Baker (1968) describes how Lindsay, Johnson, and Wallace ran for the Presidency in 1968 with none getting a majority in the Electoral College. As a deadlock developed in the House, the Senate chose R. F. Kennedy as Vice President, who then decided to suspend balloting in the House for President in order to buy time to influence their decision.

Our goal is to investigate some of the mathematical structure of the first two stages in the election of a president, and to analyze how power is distributed between the different states and the individual citizens of these different states. Our previous examples have given reason to believe that such investigations of voting influence can give rise to irregular, surprising and unpredictable comparisons in terms of a voter's ability to affect the outcome of an election. This quantitative approach will confirm the intuitive and historical evidence to the effect that there is a bias in favor of the large states and their citizens.

There is an excellent source book by Peirce (1968) on the Electoral College which discusses its history, past changes, past and present reform efforts, and potential alternatives. The appendix of this book contains a host of data relevant to any study of this institution.

7.2. The National Game. The Electoral College can be viewed as a 51-person weighted majority game between the states. This assumes that each state (plus the District of Columbia) acts as an independent player, i.e., as a free agent without prior commitments and outside influences. For this game, the sum w of the 51 weights is 538 and the simple majority quota q is 270. The weight w_i for each individual state i is equal to its number of electoral votes, and these are listed in Table 7.2.1. We will treat the District of Columbia throughout as though it were a state. These weights can change every ten years as a result of the national census, and they have occasionally been modified slightly by Congress. So the Table lists the weights as they were in 1964 and 1968 elections under the column "1961," and the way they were for the

1972 election under "1972." These are based upon the 1960 and 1970 census figures, respectively, along with the method used to determine the number of seats each state receives in the Congress.

We are interested in determining the voting power for each state in this national game. These results are given in Table 7.2.2. The first column lists in descending order the number of electoral votes, or weight w_i, which a state may have; and columns (A) and (B) list the number of such states with the corresponding weights for the years 1961 and 1972, respectively. The power indices, in the sense of Shapley-Shubik, are given in columns (C) and (D) for the two periods being considered. The first of these (C), as well as results for some earlier periods, were reported by Mann and Shapley (1962); and the second one (D) was given in Boyce and Cross (1973). Columns (E) and (F) rescale the previous two columns by multiplying by the total weight 538 so that they can be easily compared with the number of electoral votes listed in the first column. Columns (G) and (H) express these comparisons as ratios. It should be noted that the numbers in the table were obtained by decimal calculations on the computer and are rounded off, and are excellent approximations to the exact values which are in fact rational numbers since they have a factor of $51!$ in the denominator. These results indicate a clear and systematic bias in favor of the large states. However, the total magnitude of the bias is less than ten per cent, and is thus rather small in the light of the many nonquantifiable aspects in such political institutions. So, to a rather good degree of approximation, a state's voting power is proportional to its influence as expressed by its number of electoral votes.

One can design several interesting major classroom projects which are based upon or motivated by the above results, and a few of these are mentioned in what follows.

(i) Determine the power, in the sense of φ or β, for each of the 13 states in the original Electoral College. Power at other historical periods may also be of interest, e.g., at the time of the Civil War.

(ii) One can attempt to reproduce the results in Table 7.2.2 for the year 1972. This would make use of some computational techniques such as those described in Section 6.1 or 6.3, plus the design of a computer program to implement these. A detailed algorithm for doing this is described by Boyce and Cross, but one would still have to program this for his local computing facilities.

STATE ELECTORAL VOTES

State	1961	1972
Alabama	10	9
Alaska	3	3
Arizona	5	6
Arkansas	6	6
California	40	45
Colorado	6	7
Connecticut	8	9*
Delaware	3	3
Dist. of Columbia	3	3
Florida	14	17
Georgia	12	12
Hawaii	4	4
Idaho	4	4
Illinois	26	26
Indiana	13	13
Iowa	9	8
Kansas	7	7
Kentucky	9	9
Louisiana	10	10
Maine	4	4
Maryland	10	10
Massachusetts	14	14
Michigan	21	21
Minnesota	10	10
Mississippi	7	7
Missouri	12	12
Montana	4	4
Nebraska	5	5
Nevada	3	3
New Hampshire	4	4
New Jersey	17	17
New Mexico	4	4
New York	43	41
North Carolina	13	13
North Dakota	4	3
Ohio	26	25
Oklahoma	8	7*
Oregon	6	6
Pennsylvania	29	27
Rhode Island	4	4
South Carolina	8	8
South Dakota	4	4
Tennessee	11	10
Texas	25	26
Utah	4	4
Vermont	3	3
Virginia	12	12
Washington	9	9
West Virginia	7	6
Wisconsin	12	11
Wyoming	3	3
TOTAL	538	538

TABLE 7.2.1

---

* More recent results give 8 votes to Connecticut and 8 votes to Oklahoma due to a change in how some overseas population is allocated to their home state.

ELECTORAL COLLEGE POWER RATIOS

Column:	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Number of Electoral Votes w_i	Number of States i		Power Indices φ_i		Rescaled Indices $538\ \varphi_i$		Power Ratios $538\ \varphi_i \div w_i$	
	1961	1972	1961	1972	1961	1972	1961	1972
45	0	1	--	.08830938	--	47.510446	--	1.0557877
43	1	0	.08406425	--	45.226568	--	1.0517806	--
41	0	1	--	.07972734	--	42.893309	--	1.0461782
40	1	0	.07767063	--	41.786804	--	1.0446701	--
29	1	0	.05500222	--	29.591194	--	1.0203860	--
27	0	1	--	.05096311	--	27.418153	--	1.0154873
26	2	2	.04901260	.04897682	26.368783	26.349529	1.0141839	1.0134433
25	1	1	.04703309	.04699893	25.303807	25.285424	1.0121523	1.0114169
21	1	1	.03919718	.03916926	21.088085	21.073062	1.0041945	1.0034790
17	1	2	.03148765	.03146563	16.940359	16.928509	.9964917	.99579468
14	2	1	.02578474	.02576694	13.872193	13.862614	.9908709	.99018686
13	2	2	.02389839	.02388196	12.857334	12.848494	.9890257	.98834576
12	4	3	.02201920	.02200413	11.846331	11.838222	.9871943	.98651844
11	1	1	.02014710	.02013337	10.839140	10.831753	.9853764	.98470464
10	4	4	.01828200	.01826959	9.835718	9.829039	.9835718	.98290416
9	3	4*	.01642383	.01641273	8.836024	8.830049	.9817804	.98111659
8	3	2*	.01457252	.01456270	7.840015	7.834733	.9800019	.97934191
7	3	4*	.01272798	.01271944	6.847652	6.843059	.9782360	.97757982
6	3	4	.01089014	.01088284	5.858896	5.854968	.9764827	.97582844
5	2	1	.00905894	.00905301	4.873709	4.870519	.9747417	.97410373
4	10	9	.00723429	.00722957	3.892051	3.889509	.9730128	.97237690
3	6	7	.00541615	.00541245	2.913883	2.911898	.9712958	.97063223
TOTALS 538	51	51	.99999947	.9999997	537.999790	537.999803	50.2431662	50.2102844

TABLE 7.2.2

(iii) The data in Table 7.2.2 can be analyzed graphically for the 1972 results. One can plot the points (w_i, φ_i) in the plane, find the best fitting straight line, parabola, etc., and test these for goodness of fit. Mann and Shapley (1962) illustrate the data graphically for their 1961 results, and report that it has an excellent parabolic fit. Additional statistical analyses of various deviations in the data may be undertaken.

(iv) One can argue that the states are not independent agents in this national game, but instead that certain states with closely related interests or voting histories combine together and tend to vote alike. One can investigate past voting behavior of states, and aggregate together those who have voted the same way and assume that they are likely to continue to do so in the near future. These new coalitions of states can be treated as the individual players in a new version of the Electoral College game which will now have fewer players but with usually larger weights. This project can be modeled after the recent work by Boyce (1973), and there are a host of statistical hypotheses which can be assumed and investigated. For example, Boyce studies the impact of a third party movement such as the one by George C. Wallace in 1968.

(v) An ambitious project would be to investigate the procedures used by the major parties to nominate their presidential candidates at the national conventions, under say the somewhat unrealistic assumption that all delegates from a given state vote for the same candidate, i.e., the state is bound by the unit rule. This idealized game is somewhat similar to the Electoral College but the weights and number of states are determined in a slightly different fashion. The reference by David (1960) plus some more recent data may prove useful.

7.3. The State Game. Consider the game which is played in a particular state. Here the players are the individual voters in the state and a simple majority, or often just a plurality, determines the winner, who will then receive all of this state's votes in the national game discussed in the previous section. It is clear that each individual voter in a given state has equal voting power in his state game, and that his power index is 1/N, where N is the total number of voters in his state. In the following we will take N to be the total population of the state and thus assume that such N are proportional to the numbers of voters which the different states have. The number of electoral votes which a state has is based on such census figures. We also avoid consideration here of the fact that various numbers of voters may not exercise their right to vote. More important than a player's power index however, is the chance that he has of being the marginal voter in his state. Conventional wisdom often holds that this probability varies inversely with the population. On the other hand, we saw from the analysis in Section 6.2 that the possibility of being marginal, as determined by Banzhaf, is instead proportional to $1/\sqrt{N}$. This relationship was also discussed in Section 5.3 in connection with representation from multi-member districts. For example, the 1960 populations of New York State and the District of Columbia were 16,782,304 and 763,956, respectively. The chance that a voter from New York can sway his state's election compared to that for a voter from the District of Columbia is $1/\sqrt{16,782,304}$ to $1/\sqrt{763,956}$, or $(4097)^{-1}$ to $(874)^{-1}$, which gives the ratio 0.2133. In other words a voter from D. C. has 4.687 times the chance of being marginal in his "state" as the voter from New York. (i) In light of this square root factor discussed above, what possible types of errors could be introduced into the various analyses in this paper if one uses a state's population figure rather than the number of its citizens who are actually likely to vote?

7.4. The Combined Game. To obtain some measure of the individual voter's total influence in electing a president, one must take into account both the national game and the state games discussed in the last two sections. To better understand this existing system, one must somehow compose the two previous results. We have seen that it is false to presume that each citizen's relative share of power in this two level game is directly proportional to his state's weight w_i divided by its population N_i. Instead, one can combine the Shapley-Shubik index φ for the national game given in Table 7.2.2 with the Banzhaf inverse square root formula for comparing the different state games in order to get some approximate idea of the relative influence of voters from different states. These two measures reflect the power that a given state and an individual within the state have in an election. Consider, for example, a comparison between New York State and the District of Columbia using the 1961 figures in Table 7.2.2. Clearly, N.Y. has more power in the national game, and this can be indicated by the ratio of $\varphi_{NY} = .084064$ to

φ_{DC} = .005416 which is 15.521, and is a little higher than ratio of weights, i.e., $w_{NY} = 43$ to $w_{DC} = 3$ or 14.333. On the other hand, we saw in the previous section that a voter in N.Y. has only .2133 times as much chance of swaying his larger state as a voter from D.C. has of switching his district. Combining these two effects gives $(15.521) \times (.2133) = 3.311$, which indicates the relative power of an individual in N.Y. to one in D.C. in the two level election process. Note that we have merely multiplied the results of two separate games here, and that we have taken the φ measure in one case and used the Banzhaf analysis for the state games. We have not actually solved the single composed game as we did for one simple example in Section 5.3, because this composition would be a huge game in which some 100,000,000 players participate, and because the behavior of power indices (or more general values in game theory) under various compositions is not completely known. We return to this point in (iii) below.

Banzhaf (1968, Table I) has compared the ability of a citizen in one state to affect the outcome of a presidential election with the ability of one from another state, and these are shown in column (I) in Table 7.4.1. Note that N.Y. and D.C. are the extreme cases. Note also that some states have little more power than D.C., and so the latter is not the only one relatively poor in power. The low value for D.C. is expected since its $w_{DC} = 3$ is assigned in a manner different from the states and is not based so much on population. It is clear from the Table that the influence in the various state games greatly exaggerates the small bias in the national game in favor of the larger states. These results do not depend upon some quirk in our definitions of power, but are inherent in the design of the system itself when viewed as an abstract mathematical entity or structure. They surely give rise to some question about whether the "one man-one vote" principle is satisfied in presidential elections. A similar type of conclusion is reached in Brams and Davis (1974).

Some projects parallel to Banzhaf's work can be undertaken.

(i) One can use the 1972 values for w_i and φ_i given in Table 7.2.2 along with the 1970 census figures for N_i to obtain an up-dated version of the results in column (I) of our Table 7.4.1.

(ii) One can investigate various statistical deviations indicated in column (I), or its up-dated version mentioned in (i); and answer questions about what per cent of the states or people benefit or are disadvantaged by the Electoral College system. Some results of this type are given in Banzhaf (1968, Table I).

RELATIVE VOTING POWER - TABLE 7.4.1

Column: State	Population 1960 Census	I Present Plan	II Proportional Plan	III District Plan
Alabama	3266740.	1.632	1.203	1.302
Alaska	226167.	1.838	5.212	3.075
Arizona	1302161.	1.281	1.509	1.594
Arkansas	1786272.	1.315	1.320	1.459
California	15717204.	3.162	1.000	1.004
Colorado	1753947.	1.327	1.344	1.472
Connecticut	2535234.	1.477	1.240	1.362
Delaware	446292.	1.308	2.641	2.189
Dist. of Columbia	763956.	1.000	1.543	1.673
Florida	4951560.	1.870	1.111	1.197
Georgia	3943116.	1.789	1.196	1.267
Hawaii	632772.	1.468	2.484	2.092
Idaho	667191.	1.429	2.356	2.038
Illinois	10081158.	2.491	1.013	1.059
Indiana	4662498.	1.786	1.096	1.200
Iowa	2757537.	1.596	1.282	1.364
Kansas	2178611.	1.392	1.263	1.399
Kentucky	3038156.	1.521	1.164	1.299
Louisiana	3257022.	1.635	1.206	1.304
Maine	969265.	1.186	1.622	1.691
Maryland	3100689.	1.675	1.267	1.337
Massachusetts	5148578.	1.834	1.068	1.174
Michigan	7823194.	2.262	1.055	1.108
Minnesota	3413864.	1.597	1.151	1.274
Mississippi	2178141.	1.392	1.263	1.399
Missouri	4319813.	1.710	1.092	1.211
Montana	674767.	1.421	2.329	2.026
Nebraska	1411330.	1.231	1.392	1.532
Nevada	285278.	1.636	4.132	2.738
New Hampshire	606921.	1.499	2.590	2.137
New Jersey	6066782.	2.063	1.101	1.162
New Mexico	951023.	1.197	1.653	1.707
New York	16782304.	3.312	1.007	1.000
North Carolina	4556155.	1.807	1.121	1.214
North Dakota	632446.	1.468	2.485	2.903
Ohio	9706397.	2.539	1.053	1.080
Oklahoma	2328284.	1.541	1.350	1.422
Oregon	1768687.	1.321	1.333	1.466
Pennsylvania	11319366.	2.638	1.007	1.043
Rhode Island	859488.	1.259	1.829	1.795
South Carolina	2382594.	1.524	1.319	1.405
South Dakota	680514.	1.415	2.310	2.018
Tennessee	3567089.	1.721	1.212	1.291
Texas	9579677.	2.452	1.025	1.070
Utah	890627.	1.237	1.765	1.764
Vermont	389881.	1.400	3.023	2.342
Virginia	3966949.	1.784	1.189	1.264
Washington	2853214.	1.569	1.239	1.341
West Virginia	1860421.	1.506	1.478	1.514
Wisconsin	3951777.	1.788	1.193	1.266
Wyoming	330066.	1.521	3.571	2.546

(iii) The method of Owen (1972, 1975) mentioned in Section 6.3 allows one to construct a nonsymmetric value theory for the quotient game that will yield the correct symmetric indices for the composed game, and it gives good approximations to the 100,000,000-player national game between all the citizens in the U. S., and it does so with simpler computational problems than those discussed in Section 6.1. Apply Owen's approach to compute the relative voting power of the citizens in the different States in the U. S. Presidential Election.

7.5. Electoral College Reform. The unique Electoral College system was created to serve as a compromise between other possibilities. Yet the difficulties and inequities inherent in it have caused continual criticism over the years. Some changes have been made in the past and many unsuccessful reforms have been attempted. Peirce (1968) has an excellent discussion on this. An observation of party campaign strategies or a historical review of close elections confirms, at least empirically, the big state bias in this election process. A good discussion of this is given in an article entitled "The Ox-Cart Way We Pick A Space-Age President" by Hamilton (1968). Some of these close historical results are also mentioned in Banzhaf (1968, footnote on pages 322 and 323) and in Davis (1970, page 135). Interest in reform tends to peak and wane, but it was particularly intense for a few years around 1968 when Wallace's candidacy raised the serious possibility of a deadlock. Some quantitative considerations of the potential power of certain blocs of states in this third party movement are discussed in Boyce (1973). Such reform efforts were ultimately defeated by a successful filibuster carried on by Southern and small state senators led by Sam Ervin, and this is reported in the Congressional Quarterly Almanac (Volume 26, 1970, pages 840-845). Paradoxically, the system was defended by many of those states which fare rather poorly according to our mathematical analyses of power. So we continue with a system which has given rise to two deadlocks (1800 and 1824) and several presidents who received less than a majority of the popular vote (1824, 1844, 1876, 1884, 1888, 1916, 1948, 1960 and 1968). Some presidents received fewer votes than one of their opponents, and a person can theoretically lose an election in which he obtains a majority of the popular vote. Opinion polls also show that the public is in favor of reform. Further discussion on reform is given in the recent book by Best (1975) and in several references in Brams and Davis (1974).

Many alternatives to the present Electoral College system have been proposed. One frequently heard suggestion is to abolish this set-up completely, and to determine the president directly from the national popular vote. This approach could give rise to some technical difficulties if, for example, we moved towards a system of many parties. But the vote of each individual voter clearly counts the same in this case. However, the plans given serious consideration by Congress recently have usually been variations in the present Electoral College. Two types of these are referred to as the proportional plans and the district plans.

In the proportional plans, a state i with population N_i maintains its electoral vote w_i, but this weight is divided among the candidates in proportion to their statewide popular vote. These numbers need not be restricted to integers. So a citizen who changes his vote causes a shift of w_i/N_i electoral votes in the Electoral College. But voters from different states still have different influences under this plan. The relative ability of individuals from different states to affect an election can be easily computed. This was done by Banzhaf (1968) using the 1960 population figures and is given in column (II) of Table 7.4.1. Since smaller states have more electoral votes per resident than the larger ones, they have more voting power under this plan.

(i) A simple project is to recompute the numbers in column (II) using the 1972 electoral vote distribution (Table 7.2.1) and the 1970 census figures.

(ii) A more ambitious project would be to investigate a variation of the proportional plan in which votes are distributed to the candidates but not in a directly proportional plan. For example, a candidate who receives 55% of a state's popular vote might be assigned 65% of its electoral vote, or one who receives as much as 65% of the popular vote may get 95% of the electoral vote. In such a system, what sort of curves (popular vote versus electoral votes) would seem appropriate? It would be of interest to pick some trial curves and study what results they would have produced in previous elections. One may wish to limit consideration to only two major candidates, or to design a scheme for multiple parties.

In the district plans for electing the president, the voters in each state elect two electors at-large plus one elector from each congressional district in the state. Each elector then votes for the candidate who obtained the most votes in his district. The inverse square root formula can be used to determine an individual's relative chance of being marginal in the vote for the two at-large statewide delegates as well as for his own congressional district elector. His power is determined by combining these two possibilities. This case was also calculated using 1960 figures by Banzhaf (1968) and is given in column (III) of Table 7.4.1. Again, the more populous states are at the disadvantage. On the other hand, the large states have, contrary to their own interests, tended to support such plans as the proportional and district one, while the smaller ones have resisted reform in this direction.

(iii) Another project is to calculate the numbers in column (III) using the 1972 data in Table 7.2.1 and the 1970 population figures.

(iv) One can investigate certain statistical deviations in columns (II) and (III), or the new results in (i) or (iii), and discuss questions about the numbers of states or individuals who benefit or are disadvantaged by these systems. A few results along these lines are mentioned in Banzhaf (1968, Tables II and III).

(v) The Behavioral Research Council of Great Barrington, Massachusetts 01230, in their Bulletin of January, 1974, has proposed a "second revised draft" (a "third" is also planned) of the Constitution of the U. S. A., which they suggest be used for study purposes in high schools and colleges. Their proposed election procedures have a weighted-voting aspect to them, and it may be of interest to investigate these in light of the power indices discussed above.

(vi) In a recent book by geographer Pearcy (1973), he suggests that the U. S. A. should be repartitioned into a country with 38 states. He suggests that river valleys, large metropolitian areas, etc., should be part of an individual state rather than serve as boundaries for different states. His states would be more equal in size, and his plan would in theory give significant financial savings. Consider population figures, and design an Electoral College for this new U. S. A. of 38 states. His plan would also change the apportionment of Congress, another type of voting problem mentioned in Section 8.2.

(vii) In an editorial in the New York Times by Joseph Farkas on March 29, 1974, he suggests: "The principle of one man-one vote should therefore be replaced by a system of proportional representation that would weigh each man's vote in proportion to his demonstrated capability to make intelligent choices." Related letters to the editor appeared on April 11. Consider how one might implement this elitist form of democracy in light of the power indices for weighted voting.

8. ADDITIONAL VOTING PROBLEMS

8.0. Introduction. Many types of voting situations give rise to problems with a significant mathematical structure. There are several directions in which the weighted voting discussed in this paper can be extended, as well as many other sorts of voting problems which make use of other mathematical methods. A few of these extensions and alternate approaches will be mentioned in this Section. These are included in order to point the interested reader in a few of the possible directions, and no attempt is made to give any details or to cover all of the possible voting areas. For example, problems involving probabilistic, statistical, clustering, or simulation techniques are not discussed here.

8.1. Affinity of Coalitions. Experience from the real-world indicates that some voting combinations or coalitions of voters are more likely to form than others. This may be caused by personal relations between the players, by closeness of political philosophies, by other ideological grounds, or by common goals or geographical proximity. Owen (1971) has developed a modification of the Shapley value which considers this affinity of certain coalitions. The relative position of the voters is described by positioning them on a hypersphere, and then the likelihood of coalitions depends upon these locations. Owen applied this value to the eleven-party Israeli Knesset in 1965, where the parties were spread from left to right on the political spectrum. It would be an interesting project to apply this to the Israeli Knesset as it exists in 1974 (see Section 5.4 (iii)), where one would perhaps wish to collect opinions on the orthodox-nonorthodox spread as well as the liberal-conservative spectrum; or to the Danish Parliament in 1974 where some suggested data, taken from four Danish professors, is given in Section 5.4 (iv). The computation of Owen's index is usually not extremely difficult. Another view for incorporating "prior position" into the value or power index is given on pages 76-78 in Shapley (1973).

8.2. Apportionment of Congress. An important problem in the U. S. A. and elsewhere, and one with a very interesting history, is how to determine the number of seats which any state will have in the U. S. Congress. It is desirable that any such scheme be fair in the sense that it possesses a certain "consistency" property, that it gives each one of the states nearly its "quota" and that the method has a certain "monotonicity property" (i.e., avoids the "Alabama paradox"). A new Quota Method, which is the unique method satisfying such desirable properties, has just been proposed in two very interesting papers by Balinski and Young (1974, 1975). Many excellent class projects can be designed around this type of problem.

8.3. Redistricting. Another important political problem is how to partition a region into a certain number of "similar" districts, e.g., dividing a state into Congressional districts. It is desirable, or a legal necessity, that the resulting districts have certain properties, e.g., roughly equal population, contiguity (say convex), compactness (say fat or circular or square-like rather than thin), etc. One wishes to avoid gerrymandering. Several mathematical methods have been applied to this problem. Garfinkel and Nemhauser (1970) developed an algorithm by considering techniques from integer programming. They applied it to 26 census tracts in Sussex County, Delaware and to 7 districts and 39 counties in the State of Washington. They had difficulty with 55 counties in the State of West Virginia. Other approaches, such as simulation, have also been used on this problem. Several class projects along these lines are possible, e.g., consider how to construct a university senate with constituencies of about equal size where each senator represents a rather "common" or "homogeneous" group.

8.4. Sequential Voting Strategies. Many common voting procedures have an important strategic aspect to them. In many sequential situations, such as binary voting as is frequently done on a series of proposed amendments, the final outcome may depend upon the order in which the motions are considered. The result may depend upon the information about another's intentions, and thus one may benefit by deception, misleading or diversionary tactics in the earlier stages when there are multiple votes in a sequence. Optimal voting strategies and various equilibria concepts have been studied, e.g., by using the theory of games in extensive (tree) or normal (matrix) forms. Good introductions to some of these situations appear in the excellent little book by Farquharson (1969) (including historical examples in his Appendix), and in Chapter 14 (especially Section 14.8) in Luce and Raiffa (1957).

8.5. Maps. There has been recent interest in various types of geographical or political maps, called cartograms, which are not drawn in accord with standard map projections. For example, pilots use maps which have a magnification of airport areas, and one could produce highway maps in which cities or complicated interchanges are embedded "continuously" but in a much larger aerial scale. A related theorem in differential geometry appears in Sen (1975); and there are potential applications to areas in economics and political science, e.g., see the paper by Tobler (pages 215-220) in Papayanopoulos (1973).

9. LIMITS OF THE MATHEMATICAL MODELS

9.1. General Remarks. An important step in the complete model building process is to validate or verify the accuracy and usefulness of the model to the real-world situation. It is not sufficient to work in a mathematical vacuum, for one is required to check whether his conclusions are realistic or whether they must be carefully limited. Awareness of the theoretical or hypothetical nature of the results must be clearly understood when one makes use of his results in practice. This is particularly true in building mathematical models in political science, where it is so easy to be excessively isolated, to discard too many essentials, or to overlook many less quantifiable customs or procedures. In the social sciences in general it is most difficult to form complete models which contain all relevant aspects of current reality. After all, the social sciences make a much greater use of adjectives than do the physical sciences or mathematics. In this realm it is usually foolish to claim that one has obtained complete understanding, proofs, full explanation, true meaning, etc. Before employing one's abstract analysis as a true imitation of reality, one normally has to become heavily involved with complicated data and other more empirical activities.

9.2. Additional Considerations. It is clear that our mathematical analyses of certain voting situations is in many ways superficial. It does not include much of the political actuality in the global realm of power, influence, legislation, or elections in our governmental institutions. In the realities of political life there are many alliances, differences, cohesive blocs, discriminations, partisan actions, and additional sources which influence outcomes. One must consider the effects of working behind-the-scene, capabilities in caucus or in the cloakrooms, or other "secondary" considerations; as well as the superimposed or nonelected operating structures or individuals, such as a committee system with a seniority rule of succession to a strong chairman's position. One must also consider the influences of party loyalty, prestige, publicity, bossism, solidarity, persuasion, personal gains, lobbying, fractionalism, bribes, gratuities, partonage, pork barrels, horse trading, campaign financing, and so on. Hopefully, these factors are counter balanced by or at least not completely divorced from such elements as honest debate, sincere deliberations, morality, conscience, ethics, reason, fairness, equity or other rules or "laws" in the game of politics. In extreme cases the whole political structure itself is altered more drastically under coercion, threats, force, wars, coups or assassinations.

In particular, in many of our examples and suggested projects the assumption that there exists irrevocable blocs of voters is frequently less than realistic. Our applications often discounted effects due to absences, abstentions, tie breaking rules, population figures for numbers of voters, and ignored other details of voting procedures. The analysis of the Electoral College in Section 7 clearly overlooks many free-for-all and less-readily-apparent aspects of an actual national convention or political campaign. Some detailed criticisms of Banzhaf's (1968) study by a few U. S. Senators and other experts follow his paper, and also appear in the following volume of this same journal (Volume 14, pages 86-96). However, there is some historical and empirical evidence to support the big-state bias in the Electoral College as is discussed in Peirce (1968), Hamilton (1968), and Uslander (pages 61-76 in Papayanopoulos (1973)). There is clearly a great deal of material which can be used for classroom discussion or criticism of the previous projects.

At another level of criticism there is the major question of what is the role of an elected official. Should he serve in the more routine, mechanical or plebiscitarian role of the true delegate or representative, who polls, surveys or in some other way has his pulse on the will of his citizens, and who merely funnels or transmits this popular opinion to the legislature? Or should he act in a more patrician or Burkean mode where he is more of a free agent or trustee, to whom the voters abdicate power or whose judgment they rely upon? For example, U. S. Congressmen tend to act in the latter mode on questions about foreign policy or impeachment, whereas they may follow the former role in more local or domestic issues, since their reelection may so depend upon it. Additional moral and philosophical concerns about the ends, means, causes, goals and consequences of power are discussed in some of the writings by Riker.

9.3. Usefulness. It is clear that our weighted voting analysis is only a rather singular, perhaps even trivial, part of the political systems being analyzed. In many cases the weights are not the critical factors, and other practical considerations prove more important. Our conclusions admittedly fall short to the extent that the models were not general enough. However, any model must compromise between excessive complication and incorporation of the total picture. However, one should distinguish between the faults of the analyses, such as erroneous assumptions, incompleteness or incorrect conclusions, on the one hand, as against faults inherent in the actual design or very fabric of the system being modeled, on the other hand, e.g., irrational or illogical behavior, false intuition, or nonlogical structures.

One could nevertheless argue that the power indices are a worthwhile first step in a quantitative investigation of such voting situations. Although limited in scope, such analysis is hardly entirely wanting or completely irrelevant, and there should be little objection to utilizing the results in the limited or spotty context to which they apply. Such studies should be useful in setting up norms, standards, ground rules, or minimal requirements for a voting situation; and they have recently found some implementation and acceptance by the courts in practice. Although fairness or justice are hardly to be found entirely in some simple mathematical formula or symmetry, it would appear that one should attempt to avoid numerical bias, as well as other types, whenever possible. Hopefully, the power indices surveyed in this paper, will have some value as a quantitative aid in judging equality in certain political structures, and they will become part of the conventional wisdom on the subject. Knowledge of power indices might be compared to knowing the odds of obtaining various poker hands; it is rather helpful, but not sufficient in itself to being a good poker player. It is also most likely that once mathematical techniques have successfully entered into political science, they are unlikely to be completely discarded in the future; and some recent authors even go so far as to suggest a new field called "governmetrics" analogous to "econometrics."

REFERENCES

Aumann, R. J. and Shapley, L. S. *Values of Non-Atomic Games*. Princeton University Press, Princeton, NJ, 1974.

Baker, Russell. *Our Next President*. Atheneum, New York, 1968.

Balinski, M. L. and Young, H. P. A New Method for Congressional Apportionment, *Proceedings of the National Academy of Sciences*, vol. 71, November, 1974, pp. 4602-4606.

Balinski, M. L. and Young, H. P. The Quota Method of Apportionment, *The American Mathematical Monthly*, vol. 82, August-September, 1975.

Banzhaf, John F., III. Weighted Voting Doesn't Work: A Mathematical Analysis, *Rutgers Law Review*, vol. 19, 1965, pp. 317-343.

Banzhaf, John F., III. Multi-Member Electoral Districts--Do They Violate the "One Man--One Vote" Principle, The Yale Law Journal, vol. 75, 1966, pp. 1309-1338.

Banzhaf, John F., III. One Man, 3.312 Votes: A Mathematical Analysis of the Electoral College, *Villanova Law Review*, vol. 13, Winter, 1968, pp. 304-332. Also see comments by Editor (p. 303), Birch Bayh (pp. 333-335), Karl E. Mundt (pp. 336-337), John J. Sparkman (pp. 338-341), Neal R. Peirce (pp. 342-346); and in vol. 14, Fall, 1968 by Editor (p. 86), Albert J. Rosenthal (pp. 87-91) and Robert J. Sickels (pp. 92-96).

Barrett, Carol and Newcombe, Hanna. Weighted Voting in International Organizations, *Peace Research Reviews*, vol. 2, Canadian Peace Research Institute, Oakville, Ontario, Canada.

Bell, Roderick; Edwards, David V.; Wagner, R. Harrison, editors. *Political Power: A Reader in Theory and Research*. The Free Press, New York, 1969.

Best, Judith. *The Case Against Direct Election of the President*. Cornell University Press, Ithaca, NY, 1975.

Bickel, Alexander M. *Reform and Continuity: The Electoral College, the Convention, and the Party System*. Harper and Row, New York, 1971.

Boyce, William M. A "Voting Power" Analysis of Recent Coalitions in the Electoral College, mimeographed paper, 1973.

Boyce, William M. and Cross M-J. An Algorithm for the Shapley-Shubik Voting Power Index for Weighted Voting, mimeographed paper, 1973, to appear.

Bradt, R. N., et al. *Elementary Mathematics of Sets with Applications* (formerly *Universal Mathematics, Part II: Structures in Sets*). Mathematical Association of America, Committee on the Undergraduate Program, Tulane University, New Orleans, LA, 1955 and University of Buffalo, Buffalo, NY, 1959; especially Chapter 4.

Brams, Steven J. Game Theory and Politics. The Free Press, A Division of Macmillan Publishing Co., Inc., New York, 1975.

Brams, S. J. and Affuso, Paul J. Power and Size: A New Paradox, mimeographed paper presented in April, 1975.

Brams, S. J. and Davis, Morton D. Resource-Allocation Models in Presidential Campaigning: Implications for Democratic Representation, in Papayanopoulos (1973), pp. 105-123.

Brams, S. J. and Davis, M. D. The 3/2's Rule in Presidential Campaigning, The American Political Science Review, vol. 68, March, 1974, pp. 113-134.

Brams, S. J. and Papayanopoulos, Lee. Legislative Rules and Legislative Power, mimeographed paper presented in June, 1974.

Coleman, James S. Control of Collectivities and the Power of Collectivity to Act, in Social Choice, edited by Bernhardt Lieberman, Gordon and Breach, New York, 1971a, pp. 269-300.

Coleman, J. S Loss of Power, American Sociological Review, vol. 38, February, 1971b, pp. 1-17.

David, P. T.; Goldman, R. M.; and Bain, R. C. The Politics of National Party Conventions. Brookings, Washington, DC, 1960.

Davis, Morton D. Game Theory: A Nontechnical Introduction. Basic Books, Inc., New York, 1970.

Dubey, Pradeep. On the Uniqueness of the Shapley Value, Technical Report, Applied Mathematics, Cornell University, Ithaca, NY 14853, June, 1974.

Dubey, Pradeep and Shapley, Lloyd S. Some Properties of the Banzhaf Power Index, report of The RAND Corporation, Santa Monica, to appear about late 1975.

Farquharson, Robin. Theory of Voting. Yale University Press, New Haven, 1969.

Garfinkel, R. S. and Nemhauser, G. L. Optimal Political Districting by Implicit Enumeration Techniques, Management Science, vol. 16, No. 8, April, 1970, pp. B495-B508.

Gothman, H. G. and Dougall, H. E. Corporate Financial Policy. New York, 1948, pp. 56-61.

Groenning, S.; Kelley, E. W.; Leiserson, M., editors. The Study of Coalitional Behavior. Holt, Rinehart and Winston, 1970.

Haefele, Edwin T. Representative Government and Environmental Management, The Johns Hopkins Press, Baltimore, 1973.

Hamilton, John A. The Ox-Cart Way We Pick a Space-Age President. New York Times Magazine, October 20, 1968.

Harsanyi, John C. A Bargaining Model for Cooperative n-Person Games, Contributions to theTheory of Games IV, Annals of Mathematics Studies, No. 40, edited by A. W. Tucker and R. D. Luce, Princeton University Press, Princeton, NJ, 1959.

Harsanyi, John C. A Simplified Bargaining Model for the n-Person Cooperative Game, International Economic Review, vol. 4, 1963, pp. 194-220.

Herndon, James F. and Bernd, Joseph L., editors. Mathematical Applications in Political Science, VI, University of Virginia, Charlottesville, 1972; especially pp. 79-124.

Irmie, Robert W. The Impact of Weighted Vote on Representation in Municipal Governing Bodies of New York State, in Papayanopoulos (1973), pp. 192-199.

Isbell, John R. Homogeneous Games III, Annals of Mathematics Study No. 52, Advances in Game Theory, edited by M. Dresher, L. S. Shapley and A. W. Tucker, Princeton University Press, Princeton, NJ, 1964, pp. 225-265.

Johnson, Ronald E. An Analysis of Weighted Voting as Used in Reapportionment of County Governments in New York State, Albany Law Review, vol. 34, No. 1, 1969, pp. 1-45.

Kemeny, John G.; Snell, J. Laurie; Thompson, Gerald L. Introduction to Finite Mathematics, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1957, pp. 74-78 and 108-112; Second edition, 1966, pp. 79-83 and 113-116. Also John G. Kemeny, Arthur Schleifer, Jr., J. Laurie Snell and Gerald L. Thompson, Finite Mathematics with Business Applications, 1962, pp. 68-72 and 117-120.

Krislov, Samuel. Power and Coalition in a Nine-Man Body. The American Behavioral Scientist, vol. 6, April, 1963, pp. 24-26.

Lapidot, E. The Counting Vector of a Simple Game, Proceedings of the American Math. Society, vol. 31, 1972, pp. 228-231.

Leiserson, Michael. Fractions and Coalitions in One-Party Japan: An Interpretation Based on the Theory of Games, American Political Science Review, vol. 62, 1968, pp. 770-787.

Luce, R. Duncan and Raiffa, Howard. Games and Decisions: Introduction and Critical Survey. John Wiley and Sons, Inc., New York, 1957.

Mann, Irwin and Shapley, L. S. Values of Large Games, IV: Evaluating the Electoral College by Monte Carlo Techniques, RM 2651, The RAND Corporation, Santa Monica, CA, September, 1960. Reproduced in part in Shubik (1964).

Mann, Irwin and Shapley, L. S. Value of Large Games, VI: Evaluating the Electoral College Exactly, RM-3158-PR, The RAND Corporation, May, 1962. Reprinted in part in Shubik (1964).

Miller, D. R. A Shapley Value Analysis of the Proposed Canadian Constitutional Amendment Scheme, Canadian Journal of Political Science, vol. VI, No. 1, March, 1973, pp. 140-143.

Milnor, J. W. and Shapley, L. S. Values of Large Games, II: Oceanic Games, RM-2649, The RAND Corporation, Santa Monica, CA, February, 1961.

Newcombe, Hanna. If There Had Been Weighted Voting, ..., Peace Research, vol. 5, October, 1973.

Olsen, Marvin E. Power in Societies, The Macmillan Company, New York, 1970.

Owen, Guillermo. Game Theory. W. B. Saunders Company, Philadelphia, PA, 1968, especially pp. 179-185.

Owen, Guillermo. Political Games, Naval Research Logistics Quarterly, vol. 18, No. 2, September 1971, pp. 345-355.

Owen, Guillermo. Multilinear Extensions of Games, Management Science, vol. 18, No. 5, January, 1972, pp. P64-P79.

Owen, Guillermo. Evaluation of a Presidential Election, American Political Science Review, vol. 69, September 1975a.

Owen, Guillermo. Multilinear Extensions and the Banzhaf Value, Naval Logistics Quarterly, vol. 22, December, 1975b.

Papayanopoulos, Lee, editor. Democratic Representation and Apportionment: Quantitative Methods, Measures, and Criteria, Annals of the New York Academy of Sciences, vol. 219, November 9, 1973.

Pearcy, G. Etzel. A Thrity-Eight State U. S. A. Plycon Press, 1973.

Peirce, Neal R. The People's President. Simon and Schuster, New York, 1968.

Petersen. A Statistical History of the American Presidential Elections. Frederick Ungar Publishing Company, New York, 1963.

Polsby, Nelson W. and Wildavsky, Aaron B. Presidential Elections: Strategies of American Electoral Politics, 3rd ed. Charles Scribner's Sons, New York, 1971.

Rapoport, Anatol. N-Person Game Theory: Concepts and Applications. University of Michigan Press, Ann Arbor, 1970.

Ratner, David L. The Government of Business Corporations: Critical Reflections on the Rule of "One Share, One Vote," Cornell Law Review, vol. 56, November, 1970, pp. 1-56.

Richardson, Moses. Fundamentals of Mathematics. Macmillan Company, New York, revised edition, 1958, sections 64 and 149 on pages 196 and 387; third edition, 1966, sections 70 and 155 on pages 217 and 431.

Riker, William H. A Test of the Adequacy of the Power Index, Behavioral Science, vol. 4, 1959, pp. 120-131.

Riker, William H. and Niemi, Donald. The Stability of Coalitions on Roll Calls in the House of Representatives, American Political Science Review, vol. 54, 1962, pp. 58-65.

Riker, W. H. The Theory of Political Coalitions. Yale University Press, New Haven, 1962a.

Riker, William H. Some Ambiguities in the Notion of Power, American Political Science Review, vol. 58, 1964, pp. 341-349.

Riker, William H. and Shapley, Lloyd S. Weighted Voting: A Mathematical Analysis for Instrumental Judgment, Chapter 15 in Representation: Nomos X, edited by J. Roland Pennock and John W. Chapmen, Atherton Press, Inc., 1968; also as The RAND Corporation paper P-3318, Santa Monica, CA, March, 1966.

Runyon, John H.; Verdini, Jennefer; Runyon, Sally S., editors. Source Book of American Presidential Campaign and Election Statistics: 1948-1968. Frederick Ungar Publishing Company, New York, 1971.

Schwödiauer, Gerhard. Calculation of A Priori Power Distribution for the United Nations, Research Memo., No. 24, Institut für Höhere Studien, A-1060, Vienna, Austria, July, 1968.

Sen, Ashish K. A Theorem Related to Cartograms, The American Mathematical Monthly, to appear about late 1975.

Schubert, Glendon A. Quantitative Analysis of Judicial Behavior. The Free Press, Glencoe, IL, 1959; especially Chapter IV.

Shapley, L. S. A Value for n-Person Games, in Annals of Mathematics Studies, No. 28, Contributions to the Theory of Games, vol. II, edited by H. W. Kuhn and A. W. Tucker, Princeton University Press, Princeton, NJ, 1953, pp. 307-317.

Shapley, L. S. Simple Games: An Outline of the Descriptive Theory, Behavioral Science, vol. 7, No. 1, January, 1962, pp. 59-66; also The RAND Corporation, RM-1384, Santa Monica, CA, 1954.

Shapley, L. S. and Shubik, Martin. A Method for Evaluating the Distribution of Power in a Committee System, American Political Science Review, vol. 48, September, 1954, pp. 787-792. Also Chapter 9 in Shubik (1964).

Shapley, L. S. and Shapiro, N. Z. Values of Large Games, I: A Limit Theorem, RM-2648, The RAND Corporation, Santa Monica, CA, November, 1960.

Shapley, L. S. Values of Large Games, III: A Corporation with Two Large Stockholders, RM-2650-PR, The RAND Corporation, Santa Monica, CA, December, 1961.

Shapley, L. S. Notes entitled "Political Science" on pp. 37-92 of Notes of Lectures on Mathematics in the Behavioral Sciences, notes by Henry A. Selby, MAA Summer Seminar at Williams College, Mathematical Association of America, 1973.

Shubik, Martin, editor. Game Theory and Related Approaches to Social Behavior. John Wiley and Sons, Inc., New York, 1964; especially Part 3: Political Choice, Power, and Voting.

Spilerman, Seymour and Dickins, David. Who Will Gain and Who Will Lose Influence under Different Electoral Rules? American Journal of Sociology, vol. 80, September, 1974, pp. 443-477.

Taylor, Michael. Proof of a Theorem on Majority Rule, Behavioral Science, vol. 14, 1969, pp. 228-231.

von Neumann, John and Morgenstern, Oskar. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ, 1944; 2nd ed., 1947, 3rd ed. 1953.

Chapter 4
A MODEL FOR MUNICIPAL STREET SWEEPING OPERATIONS

A. Tucker*
and
L. Bodin
State University of New York
at Stony Brook

1. Introduction

An enormous variety of human activities can be the subject of mathematical modeling. To optimize the performance of an auto assembly line, the process of putting two screws into an auto chassis may be broken into, say, twenty-five carefully defined steps and then an order and timing of these steps is determined in a way that makes the process as easy, error-free, and quick as possible. In a recent criminal trial in New York, the defense attorneys relied heavily on computer-generated profiles to pick a jury most favorably disposed towards the defendants.

When an activity, such as one of those mentioned above, entails substantial costs in time, money, or people, then the precise analysis produced by a mathematical model is often invaluable. At the same time, it must be remembered that such analysis itself requires time, money, and special personnel. While most branches of the federal government and all large corporations have used mathematical models to plan their operations for many years, state and municipal operations are just now beginning to make use of this aid. Recently, New York City has taken the lead in using mathematical modeling to improve urban services and cut their costs. In this module, we shall examine a model that was developed to route and schedule street sweepers.

A few words are in order to explain the special difficulties inherent in building good mathematical models for municipal services. It can take up to a year and thousands of dollars just to collect the data (which often varies seasonally) needed in the model, yet few municipal governments feel they can justify such long term investments when they are usually faced with immediate and more pressing problems in the community. Often the task to be modeled is performed by workers who belong to a strong union, and the union leaders will take the position that they will not let any so-called experts in City Hall tell their men in the streets how to perform their job. On the mathematical side, a major difficulty is that mathematical models need a precise formulation of the problem and its constraints. But there are many constraints in urban problems that are hard to quantify, involving administration (i.e., bureaucracy), politics, unions and the men who perform the service, plus many practical difficulties of which, as the unions rightly claim, only the men on the streets are aware. Moreover, the more precisely one does define the problem, the more unwieldy it usually is to solve. Thus to make a nice mathematical solution possible, one often has to make idealized assumptions. Yet the results that come out of the mathematical analysis must be realistic and workable. For example, when things go wrong or special events occur, a district dispatcher should still be able to get the job done as well as he could with the old ways.

---

* The work of the first author on this module was partially supported by National Science Foundation Grant GP-P033568-X00.

It should be noted that some attempts at optimization of services have been in practice for many years. For example, garbage trucks do not just go out into an area and go down streets however they wish. The routes have been planned by capable people who looked carefully at maps of the areas to be covered. One of the first objectives in any computer analysis is to see how closely the hand-drawn routes approximate the optimized mathematical solutions.

We have chosen to discuss the problem of efficient street sweeping because its analysis contains all the major features typical of a large-scale routing or scheduling problem, yet the mathematical techniques required are comparatively simple and for small problems, the calculations can be done by hand. As in many routing problems, some parts of the sweeping problem can be handled very precisely, while other parts require simplifying idealizations and even guess work. This problem also has its share of extra administrative and practical constraints.

The model described here was developed for the New York City Department of Sanitation by L. Bodin and S. Kursh in the Urban Science Program at the State University of New York at Stony Brook. The New York City Sanitation Department has a $200,000,000 annual budget of which $10,000,000 goes to street sweeping. The computerized sweeper routing based on this model, if implemented city-wide, is expected to save close to $1,000,000. The model was used in a part of the District of Columbia where it cut costs by over 20%.

Section 2 describes our sweeping problem and Section 3 presents the basic mathematical model. Section 4 develops the analysis of the model. Section 5 gives the algorithms that are required in our analysis. Section 6 contains a summary of the steps in the analysis and Section 7 gives a detailed example. Section 8 has some comments about computer implementation of the model. Section 9 presents some final comments and extensions. Section 10 contains a set of exercises. While reading the lengthy Section 4, some readers may want to refer to the summary in Section 6 and to the example in Section 7.

For instructors: A major reason for the presentation of this particular problem is the natural way it lends itself to an open-ended teaching style, in which the students themselves can develop, with the guidance of an instructor, the mathematical model and an analysis of the problem in terms of this model. Because there are several basically independent steps in this analysis, it is also possible to present some of the steps in a lecture and leave others to the class. Two of the three minimization algorithms required in the solution, the shortest path and minimal spanning-tree algorithms, could be "discovered" by the class.

If an open ended approach is used, it is important to guide the students into the "route first - cluster second" method. Also, students should not worry about disconnected networks or the turns made at each corner until the other problems have been analyzed. Beyond this, the students can progress as they wish. Several different proofs exist for Theorems 1 and 2. (When students later are studying the problem of turns at each corner, they will discover the proof of Theorem 1 given in this text.) If students write a computer program for street sweeping, they will find several minor but challenging loose ends in the model (some are indicated in Section 8). Even though these problems often can be ignored in actual practice, some class discussion concerning heuristic solutions to these problems is worth including, if time permits.

Note that even in an open-ended approach, the instructor should consider distributing sections of the text after the students have developed the particular analysis.

If the material is presented in a normal lecture method, it is advisable to insert examples after each stage of the analysis in Section 4. In addition to the extended example in Section 7, several of the exercises in Section 10 can serve as examples. The instructor may also choose to introduce the algorithms when needed rather than waiting until the analysis is completed. With students who have a limited exposure to graphs, the instructor should consider introducing other types of problems that are naturally modeled with graphs. The exercises have several examples of graph modeling, and more may be found in references [1, 2].

2. Statement of the Problem

Our problem is to design an efficient set of routes for sweeping the streets in some city. While we speak of "sweeping a street," it is actually only the sides of a street, along the curb, that are to be swept. This task is performed by vehicles called "mechanical brooms," or brooms, for short.

At first the problem seems to be trivial: send a broom up and down the length of one street, north-south or east-west, and have it repeat this on the next parallel street; and partition the city up into sets of north-south and east-west streets such that each set can be covered during the time of a single broom's period on the streets (two to four hours, depending on union work rules). However, to sweep along a curb there must be no parked cars present. Thus we quickly see that we are at the mercy of parking regulations. Many smaller cities with well-disciplined citizenry can institute parking regulations designed to coincide with the type of broom routes mentioned above. However in a city like New York City, simple consistent parking regulations,* such as allowing cars to park on alternate sides of the street on alternate days, are essential in most business districts. In addition, major arterial routes cannot be swept during rush hours. In residential and manufacturing districts, where the full parking capacity of the streets is needed much of the time, special regulations are needed. One-way streets and areas where streets do not have a common north-south and east-west alignment cause further complications. Finally, a city is usually divided up into administrative districts and broom routes may not cross district lines.

---

* Even if one could make a good case for changing certain regulations, the changeover cost might be prohibitive. It costs about $2000 per mile to install new parking signs, and about $1000 per mile to change them. New York City has 11,000 miles of streets.

As we progressed through the preceding paragraph, the character of our problem changed from trivial to extremely complex. We now seem to be overwhelmed by a large and disparate set of constraints.

To close this section, we note one additional constraint. When among one-way streets, broom routes should try to avoid turns where the curb to be swept will switch from the left to right side of the vehicle--or vice versa (this requires the driver to get out and reposition the brooms on the other side of the vehicle). To a lesser extent, they should also avoid U-turns or left turns.

3. The Mathematical Model

The sweeping problem's diverse set of constraints all but force us to seek an abstract model. We need a general structure that can incorporate all this information without getting bogged down by its complexity. This structure is a directed graph. A *directed graph* $G = (N, E)$ consists of a set of N of *nodes* and a set E of edges, each directed from one node to another node. References [1, 2, 3] discuss the theory of graphs and their applications. (In the next section we also encounter undirected graphs, whose edges simply link pairs of nodes without direction.) An edge represents *one side* of a street (in one block). See Figure 1. In this problem we will assign a "length" to each edge, namely, the length of time it takes a mechanical broom to traverse the edge. In Figure 1b, the lengths are written beside each edge. A node will usually represent a street corner. An edge e_i from node x to node y can be written as $e_j = (x,y)$. Indeed, it turns out to be most convenient to represent a graph as a set of nodes and a set of directed connections between pairs of nodes. (The connection (x,y) would be listed twice if there is a one-way street from x to y.) Hence our problem about edges

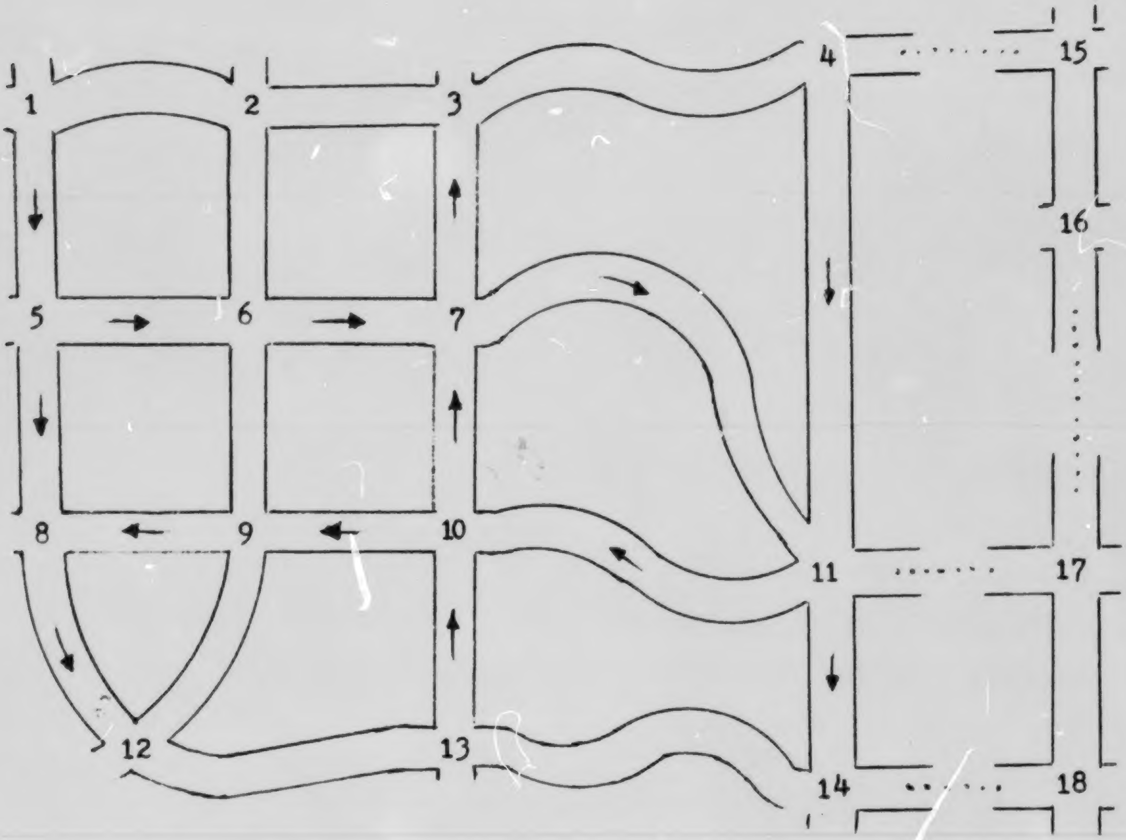

4
15
1
2
3
16
5
6
7
8
9
10
11
17
12
13
14
18

Figure 1a: Streets in a district; streets without arrows are two-way.

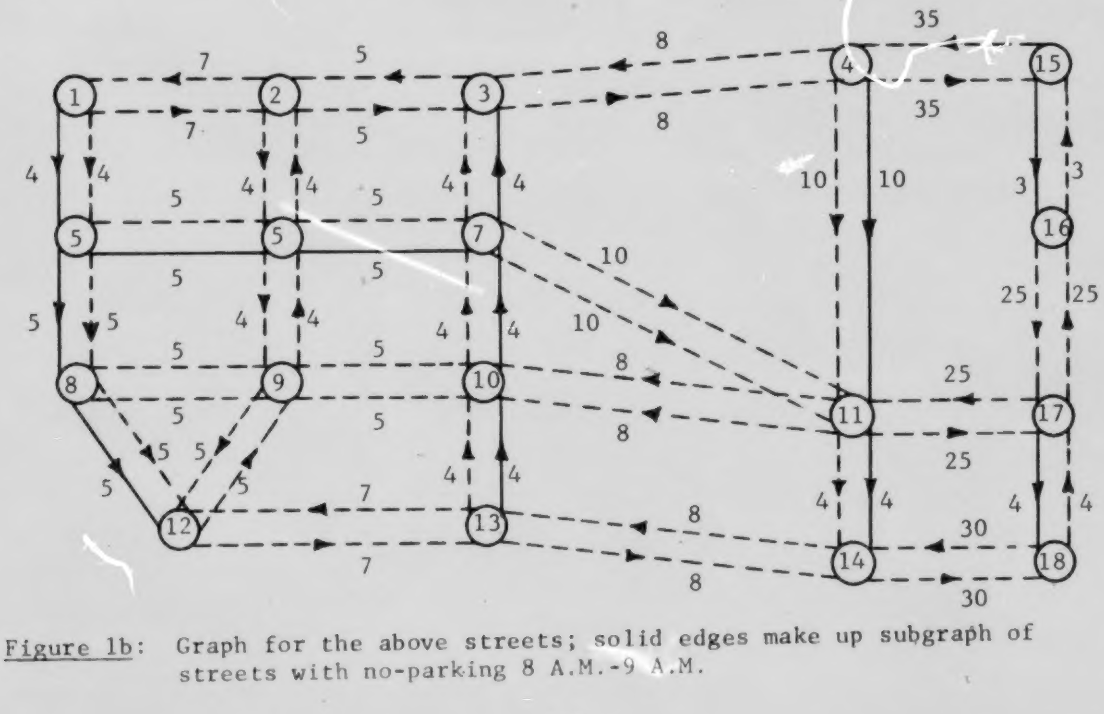

35
8
4
15
7
5
1
2
3
35
7
5
8
4
4
4
4
4
4
10
10
3
3
5
5
5
7
16
5
5
10
25
25
5
5
4
4
4
4
10
5
5
8
25
8
9
10
11
17
5
5
8
25
5
5
5
5
4
4
7
4
4
4
4
8
30
12
13
14
18
7
8
30

Figure 1b: Graph for the above streets; solid edges make up subgraph of streets with no-parking 8 A.M.-9 A.M.

will be presented in a node-based model. This is natural since it is the nodes, not the edges, that play the active role in our problem; the decision about which edge to sweep next is made at a node.

Next let us consider the time constraints imposed by parking regulations. These constraints turn out to be a blessing in disguise, for they force us to break the original large problem into many short, and manageable, problems. That is, for each period of the day or week, and for each administrative district, we form the subgraph of edges on which parking regulations (and rush-hour constraints) permit sweeping. See Figure 1. Since any given edge should not appear in more than one subgraph, we occasionally must impose artificial constrations (to determine the unique subgraph in which to put an edge that represents, say, a side of a street where parking is always banned) to narrow down the period when an edge can be swept.

Our problem can now be given a mathematical formulation: for each of the subgraphs generated above, find a minimal set of routes (i.e., minimize the number of routes) that collectively cover all the edges in the subgraph. Lengths (times for traversal) have been assigned to the edges and each route cannot require more time than the length of the no-parking period. In addition, the route should be as free of U-turns and other undesirable turns as possible. The primary difficulty arises from the fact that a street sweeper frequently must raise up its broom and travel some distance along streets that are not to be swept (or were swept already) to get to a new sweeping site. (Typically a sweeper travels twice as fast, making the effective length of streets half as long, when the broom is up.) Some students might want to see how few hour-long routes they need to sweep all the solid edges in Figure 1b.

The problem of instituting parking regulations in a district with no regulations can be done simultaneously with the minimal routing (see exercise 16). There are additional questions that we shall mention in Section 9 and in the exercises, such as linking together routes in successive, short periods.

We can now restrict our attention to the problem of finding a minimal set of routes (of a given length) covering the edges of a directed graph. The solution to this problem would then be applied to each of the subgraphs representing streets to be swept in a given district during a given period.

4. An Analysis of the Mathematical Problem

There are two different approaches to the problem of seeking a minimal, or in practice, near minimal, number of routes to cover the edges of a directed graph. We can either divide the graph up into subgraphs of a size small enough that each subgraph can be covered by a single route (of a given length), or we can find one extended route covering all the edges in the graph and then break it up into appropriate-size routes. These methods are called "cluster first-route second" and "route first-cluster second," respectively. We shall use the term "tour" for an extended route covering all edges. The route-first method is much more tractable because an exact mathematical solution exists to the minimal tour problem. However, as we see in the example in Section 7, the subsequent clustering need not lead to a minimal set of routes. On the other hand, when we try to cluster first, cutting the graph into the right subgraphs, we must rely on essentially guess work and quickly drawn (inefficient) possible routes. This causes us to cluster into conservatively small subgraphs. Another major problem is that after several "nice" subgraphs have been cut out, the remaining edges may be scattered about in many small sections. There are some other not so obvious drawbacks to the cluster first method. The net effect is that the route first approach is easier mathematically and gives better solutions for most graphs. It should be mentioned that the cluster first method does have one merit that is important to administrators. The routes obtained by that method are generally contained in a nice, compact region and different routes do not cross. This makes it easy to determine whose route contains a missed (unswept) street without need for recourse to a detailed map marking out each route. The cluster first method is discussed in [6].

We shall now develop the route first-cluster second approach. So, our task will be to obtain a minimal tour that covers all the edges of a given graph. The subsequent job of breaking up that tour into feasible routes is treated later. It is worthwhile to note how much we have been able to simplify the complex problem we had in Section 2.

It is convenient to assume that the tour must be a circuit. (Formally, a circuit is a sequence of edges $(e_1, e_2, \ldots, e_n)$ such that the end node of e_i is the start of e_{i+1} and e_n ends at the start of e_1.) This assumption means that there will be many possible ways to break up this tour later into subtours.

119

While in general any circuit (minimal or not) covering all edges will pass some edges twice, in some graphs it is possible to obtain a circuit that traverses each edge of the graph exactly once. A circuit covering all edges once (and all nodes) is called a Euler circuit (or Euler tour). Clearly, any Euler circuit is a shortest possible tour, since by definition of our problem, we must traverse each edge at least once. It turns out that the conditions for the existence of an Euler circuit will tell us how a shortest-tour construction must procede even when an Euler circuit does not exist.

There is an obvious condition for the existence of an Euler circuit. Namely, there must be as many edges directed into each node as are directed out of it, since a circuit arrives at a node as many times as it leaves a node. Let us define the inner (outer) degree of a node to be the number of edges directed into (out of) the node. A graph is connected if there exists a set of edges, disregarding their directions, that links together any given pair of nodes. Clearly, graphs with Euler circuits are connected.

Theorem 1: A directed graph has an Euler circuit if and only if the graph is connected and for each node, the inner degree equals the outer degree.

Proof: We have alread observed that graphs with Euler circuits satisfy these two conditions. Let us have a graph satisfying the two conditions. At each node, we arbitrarily pair off, and tie together, the inward and outward edges. Thus we have tied all the edges together into long strings which, having no ends, must be circuits. We can repeatedly combine any two circuits with a common node (by repairing the edges (on the two circuits) incident to the common node). By the connectedness of G, the result cannot be two or more vertex-disjoint circuits. Thus we obtain a single circuit, an Euler circuit. Q.E.D.

Note that the construction in the above proof would allow us to pair inward and outward edges at each node in a manner designed to minimize unwanted turns (U-turns, etc.) at each node.

Now let us consider the problem of finding a minimal tour covering all edges in an arbitrary connected graph. We handle disconnected graphs by solving our problem for each connected part, called component, of the graph (later in this section we describe how to tie these solutions together). The difficulty in arbitrary graphs arises at nodes whose outer degree does not equal the inner degree. At such corners, a vehicle will lift up its brooms and drive over to some other corner where the broom will be lowered and sweeping will recommence. The time spent with the broom up is called deadheading time. This is what must be minimized (the time needed to sweep is predetermined. Define the degree $d(x)$ of a node x to be the outer degree of x minus the inner degree. If a node x has $d(x) < 0$, i.e., an excess of inward degrees, then our tour must eventually deadhead on an edge when it leaves x. A similar difficulty arises if $d(x) > 0$.

We can draw a graph of the route of a tour. See the example in Figure 2. Here we represent the added deadheading edges as dashed edges. These deadheading edges are drawn from the larger graph of all streets in the city. The dashed edges are duplicates of existing (no parking) edges when the broom deadheads down a street that must be swept. In other places, the dashed edges represent sides of streets where parking is not banned. Observe that the added deadheading edges necessarily expand the original connected graph into a graph having an Euler circuit, that is, by Theorem 1, into a connected graph where each node x has $d(x) = 0$. So the problem of finding a minimal tour becomes the problem of finding a minimal-length set of edges whose addition to the original graph balances the inward and outward degrees of each node.

Theorem 2: Let G be a directed graph with a length assigned to each edge, and let H be a larger graph containing G. Let A be a minimal-length collection of edges (a single edge may be counted several times in A) drawn from the graph H such that the addition of the edges of A to G makes $d(x) = 0$ for each node x in the new graph. We assume such a set A exists. Then A may be partitioned into paths (a consecutive sequence of edges) from nodes of negative degree to nodes of positive degree. If $\deg(x) = -k$ (or $+k$) in G, then k of the paths start (end) at x.

Proof: Let G^* be the graph generated by just the edges in A (with k copies of an edge that is counted k times in A). The only nodes of G^* with non-zero degree will be the nodes of G with non-zero degrees, since the sum of the degree of a node in G and the degree in G^* is zero. Let us arbitrarily pair off, as in the proof of Theorem 1, inward and outward edges at each node of G^* as much as possible. Thus if $d(x) = -3$ in G^*, three inward edges at x would remain unpaired. We again get a set of long strings of edges. Except now, some strings have beginnings and ends. The beginnings are necessarily nodes of positive degree (negative degree in G) and the ends, nodes of negative degree (positive in G). If any string formed a circuit, we could drop that set of edges from A without changing the degrees of any nodes. So by the minimality of A, all the strings are paths from nodes of negative degree in G to nodes of positive degree in G. Q.E.D.

Theorem 2 tells us how to minimize the deadheading time. We must look at all ways of pairing off negative nodes with positive nodes with deadheading paths (a negative or positive node x would occur in $|d(x)|$ deadheading paths) and then pick the set of pairings that minimizes the total deadheading time, i.e., minimizes the sum of the lengths of the shortest paths between the paired negative and positive nodes. Recall that when deadheading, a mechanical broom normally travels two times as fast and so the total deadheading time is actually half the sum of the lengths. When the minimal set of deadheading pairings is found, we add the dashed deadheading edges to our graph. Now the resulting graph has zero-degree nodes and we can apply the method in Theorem 1. Note that when we are building an Euler tour in Theorem 1, it could happen that we create a deadheading path from the negative node x to a node of zero degree. For example, this happens in the tour in Figure 2 when the positions of e_8 and e_9 are exchanged. For once the minimal set of dashed edges has been appended, we can trace out an Euler circuit in many different ways. Any such Euler circuit will be a minimal deadheading route for a mechanical broom.

To solve the minimal pairing problem, we need a matrix giving the lengths of shortest paths between i-th negative node x_i and j-th positive node y_j, for all i, j. (Of course the routes of the shortest paths are also needed.) If the graph is disconnected, we should solve the pairing problem for the whole graph at once since an optimal pairing often links nodes in different components (this situation occurs in the example in Section 7). Note that Theorem 2 does not require G to be connected. In the next section, we give an algorithm for finding the shortest paths between pairs of nodes in a graph. Remember when looking for the shortest path, we need not restrict ourselves to the graph at hand, the graph of no-parking streets at a certain time. Rather we should look at the graph for all the streets in that district of the city. We note that for large problems, a geographical estimate can be used to find the shortest paths between the negative and positive nodes. The computer program used in the New York City project knew the coordinates of

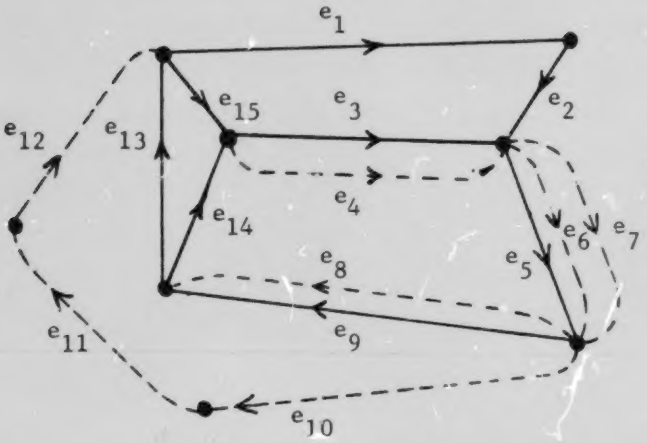


Figure 2: Directed graph of edges to be swept (solid edges) with additional deadheading edges (dashed edges) needed for a complete tour. A possible tour is $(e_1, e_2, e_5, e_9, e_{14}, e_4, e_6, e_8, e_{13}, e_{15}, e_3, e_7, e_{10}, e_{11}, e_{12})$.

	40	80	20
60	3	6	7
50	8	2	4
30	5	4	1

Figure 3: Typical data for a transportation problem; see pages

Turn	Weight
Straight ahead	0
Right turn	1
Left turn	4
U turn	8
*Switch sides of street	10 additional
Raise or lower broom	5 additional

*Only applies when sweeping

Figure 4

each node (corner), and by adding the differences of the x-coordinates and of the y-coordinates of two nodes, it derived what is often called the "Manhattan" distance. Multiplying this number by an appropriate constant yields a good estimate of the travel time between the two nodes. After a minimal pairing is found from these distances, we go back to our shortest-path algorithm to find the specific deadheading edges (and actual length) of a shortest path between each two paired nodes.

Let us now reformulate the problem of finding a minimal set of pairings between the negative and positive nodes. We have a matrix A with a row for each negative node and a column for each positive node. Entry a_{ij} is the "cost" (length of shortest path) in going from the i-th negative node x_i to the j-th positive node y_j. On the left side of the matrix beside row i, we enter the "supply" at x_i, $b_i = |d(x_i)|$, and on the top of the matrix above, column j, we enter the "demand" at y_j, $c_j = d(y_j)$. Our problem in this setting is to minimize the total cost of "moving" the supplies from the x_i's to meet the demands of the y_j's. Problems of this sort arise frequently in operations research and are called transportation problems. See the example in Figure 3, which could have come from the problem of finding a minimal-cost schedule for shipping grain from three different grain elevators, having supplies of 60, 50, and 30 tons, respectively, to three mills with demands of 40, 80, and 20, respectively. Note that we always need $\Sigma b_i = \Sigma c_j$ (if supplies exceed demand, we can add an artificial column to balance this equation). We give an algorithm for the transportation problem in the next section.

The solution of the transportation problem gives a minimal set of pairings, that is, it tells us which paths of deadheading edges to add. Let G' be G plus these additional deadheading edges. We now build an Euler circuit for G'. If G' is not connected, we build an Euler circuit for each component. Our next step is the pairing of inward and outward edges at each node needed in the construction in Theorem 1. We do this pairing in a manner aimed to minimize undesirable turns. We assign a weight to each possible type of pairing. See Figure 4. An extra weight is given to switching from dashed to undashed edges (or vice versa), because one wants to raise or lower the broom as infrequently as possible. (One shortcoming with this model is that in the final Euler circuit there is no way to force a deadheading path that starts from right-side sweeping to terminate at an edge requiring right-side sweeping. For this reason, we do not assign a side-of-street to deadheading edges. No changing-side-of-street penalty can occur with them.) At each node, we seek a pairing of the nodes's inward and outward edges in G' that minimizes the sum of the weights. To put this problem in a standard form, we make a matrix $W^{(k)}$ for the k-th node v_k with a row for each inward edge to v_k and a column for each outward edge from v_k. Entry w_{ij} in $W^{(k)}$ is the weight fro pairing the i-th inward edge with the j-th outward edge. The problem of a minimal pairing of the row elements with the column elements is called the assignment problem. The assignment problem is actually a "degenerate" transportation problem with row supplies and column demands all equal to one. In practice, a node usually has at most two inward edges and thus at most two pairings exist--so one computes the weights of each pairing and picks the lesser one. (Another possible approach to minimizing undesirable turns is to minimize the largest weight that occurs in the pairing--this is called a bottleneck problem--but this approach often gives bad pairings.)

After the assignment problem has been solved at each node, we form circuits as in the proof of Theorem 1. We combine these circuits together as in that proof. The re-pairing needed to combine pairs of circuits may occasionally form bad turns. However, in practice there are few of the circuits and so it is not important to worry about finding an optimal way to paste them together (in a typical 1000-edge street-sweeping graph, only 5 to 8 circuits occur; in small graphs, there is little, or often no, choice of where to paste).

Finally, we consider how to handle the case of a disconnected graph. Assume we have Euler circuits (forming minimal-length tours) for each component of G. We pick an arbitrary component G_1 and start the prescribed tour of G_1 at a chosen node x_0. Then at some x_1 in G_1, we cross over to some node x_2 in component G_2' and now begin the tour of G_2. If we do not leave G_2 for another component, then we complete the tour of G_2 and go from x_2 back to x_1 and resume the tour of G_1. (Note that if the tour on G' deadheads from x_1 to some node x_1', then we would do better to return directly to x_1'; such short-cuts are difficult to systematize and are ignored at this stage, but in the final stage of our procedure--on the last page of this section--obvious short-cuts can be inserted by hand.) We shall require that the tour of G_2, which may be interrupted by side tours to other components, should always be completed before we ever come back to G_1; further, once

G_2's tour is completed, we must return directly to node x_1 in G_1. See the example in Figure 5a. These requirements apply to all components.

We can define a component graph H, an undirected graph, with a node for each component of G' and an undirected edge between node x_i and x_j in H if our tour crosses between components G_i and G_j. See Figure 5. We have implicitly required that H contain no circuits. A connected graph with no circuits is called a tree (that is what it looks like). If the length of of an edge (x_i, x_j) in the component graph is the shortest distance of a round-trip (paths in both directions) between some node in G_i and some node in G_j, then an optimal H would minimize the sum of the edges. Such a tree is called a minimal spanning tree. It is usually true, and we assume thus, that a minimal spanning tree of components minimizes the amount of deadheading time needed to link together the tours of the individual components (however there are some cases where circuits in H may give an improved solution; one case occurs when the same node in G_i is the closest node of G_i to several nearby components--see the counterexample in Figure 6). In the next section, we give an algorithm for finding a minimal spanning tree in an undirected graph. To find the distance between the closest nodes in each pair of components, we again use the approximate lengths supplied by the Manhattan distance (based on the difference of the coordinates). With this measure, it is not too hard to sort through (by hand or programmed heuristics) the many possible pairs of nodes to find the closest pair of nodes for a given pair of components.

As before, once the minimal spanning tree is found and we know which pairs of nodes between different components are to be linked, then we turn to our Shortest Path Algorithm to find actual shortest paths in each direction between these pairs of nodes. Using these shortest paths, we join the tours of the individual components of G' into a tour of all of G' --a tour covering all edges of the original graph G.

The only remaining step is breaking up the grand tour into routes of a length that do not exceed the length of the no-parking period associated with the given graph. We start at an arbitrary node x_0 and follow the tour as far as we can within the time period. Suppose the first route ends at node y_0. The next route starts at y_0 and continues as far as possible. The process is repeated until the whole tour has been covered and we have returned

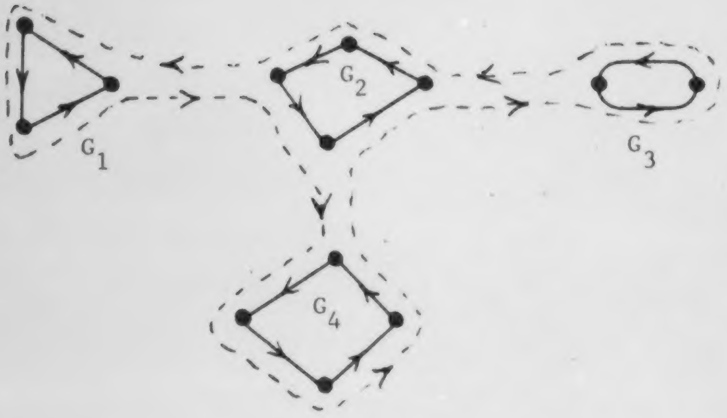


Figure 5a: Tour linking components

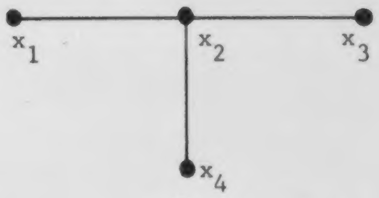


Figure 5b: Associated component graph

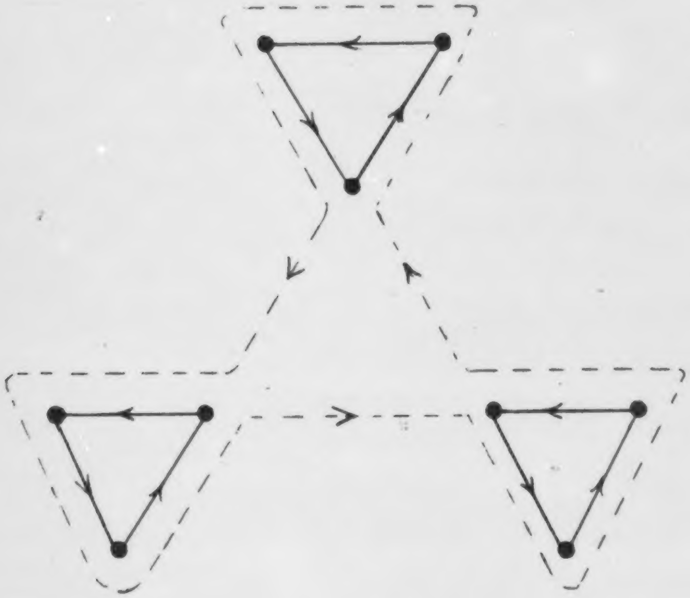


Figure 6: A graph in which the minimal tour does not form a tree among the components.

to x_o. Note that if a route started (or ended) with a stretch of deadheading time, that part of the tour can be omitted. How much deadheading we can omit depends on which node was chosen as x_o. To try other possibilities, we pick the node following x_o on the first route and let it be the starting point for generating successive routes. Then we try the node after that as the start. We repeat this process until we reach y_o, the end of the original first route. Since one of the routes must begin somewhere in the stretch between x_o and y_o, we have thus checked all possible ways of generated routes. The starting point that resulted in the fewest routes is the one we use.

It is important to note that while the above single-tour solution was optimal (except for possible short-cuts or circuits in the process of linking the components of G'), the multiple-route solution is not guaranteed to be optimal. This is because not all positive or negative nodes need be paired off in a multiple-route solution (a route could end or start at such nodes). The use of a minimal spanning tree to join components presumed that the tour had to return to the component at which it started. This is not necessary in a multiple-route solution. The example in Section 7 illustrates this difficulty. However, in connected graphs, the solutions of our model are usually quite close to optimal. In the routing system used in New York City, the (final) stage of breaking up the grand tour into routes was done by hand. In addition to finding possible improvements in disconnected graphs, hand routing allows for practical adjustments (and incorporating the possible short-cuts when linking together components of G') and permits breaking up the tour in a way that started or ended many of the routes close to the district garage.

This completes the analysis of the broom routing problem in one district during one no-parking period. Of course, in a real problem this must be repeated for all districts and all periods. In the next section, we present the three special algorithms we need to find shortest paths, to solve a transportation problem, and to find a minimal spanning tree. Section 6 has a summary of the preceding analysis.

5. Minimizing Algorithms

In the preceding section, we showed how to break the broom routing problem into several parts; we identified places where particular minimization subproblems arose; and we showed how to piece the solutions of those subproblems together to get an efficient solution to the original routing problem. In this section, we present algorithms to solve the different minimization subproblems. There were three such minimization problems and all are basic problems of operations research. Actually, a fourth problem, the assignment problem, also arose but we chose to convert it into a transportation problem. A slightly faster algorithm specially for the assignment problem also exists. All these algorithms are discussed at greater length in standard operations research texts such as [4, 5].

a. Shortest Path Algorithm

The algorithm we shall discuss is due to Dijstra and can be used to find shortest paths from any given node a to any other node z, or from the node a to all other nodes. Recall that the analysis developed in the preceding section needs such an algorithm to find shortest paths between various pairs of negative and positive nodes and between closest nodes in two neighboring components.

Shortest Path Algorithm for shortest paths from a given node a to each node. Let $k(e)$ denote the length of edge e. Let the variable m denote a "distance counter." For increasing values of m, the algorithm labels nodes whose (minimal) distance from node a is m.

1. Set $m = 1$ and label node a with $(-,0)$, where the "-" represents a blank.

2. Check each edge $e = (p,q)$ from some labeled node p to some unlabeled node q. Suppose p's labels are $(r,s(p))$. If $s(p) + k(e) = m$, label q with (p,m).

3. If all nodes are not yet labeled, increment m by 1 and go to step 2. Otherwise go to step 4. (If we are only interested in a shortest path to z, then we go to step 4 when z is labeled.)

4. For any node y, a shortest path from a to y has length $s(y)$, the second part of the label of y, and such a path may be found by backtracking from y (using the first part of the labels) as described below.

A brief discussion of the idea behind this algorithm and an example follow.

130

Observe that instead of concentrating on the distances to specific nodes, this algorithm solves the questions: how far can we get in 1 unit, how far in 2 units, in 3 units, ..., in m units, ...? Formal verification of this algorithm requires an induction proof (based on the number of labeled vertices). The key idea is that to find a shortest path from a to z we must first find shortest paths from a to the "intervening" nodes. Suppose that $x_1, x_2, \ldots, x_k$ are the set of nodes with edges going to node y, that for each x_i we found a shortest path P_i from a to x_i with length s_i, and that $k_i = k(x_i, y)$ is the length of the edge from x_i to y. Since a shortest path to y must pass through one of the x_i's, the length of the shortest path from a to y equals $\min_i (s_i + k_i)$. Moreover, if x_i is a minimizing node, then P_i followed by edge (x_i, y) is a shortest a-y path. To find this shortest path, we shall not need to obtain the distance from a to all nodes adjacent to y, since only those x_i's that are closer to a than y are of interest.

Let $P_n = (x_1, x_2, \ldots, x_n)$ be a shortest path from x_1 to x_n. Then $P_n = P_{n-1} + (x_{n-1}, x_n)$, where $P_{n-1} = (x_1, x_2, \ldots x_{n-1})$ is a shortest path to x_{n-1}. Similarly $P_{n-1} = P_{n-2} + (x_{n-2}, x_{n-1})$ and so on. Then to record a shortest path P_i to x_i, all we need to store (as the first part of a label in the above algorithm) is the name of the next-to-last node on the path; namely, x_{i-1}. To get the node preceding x_{i-1} on P_i, we go to the first stored node for P_{i-1}, that is, the next-to-last node on P_{i-1} which is x_{i-2}. By continuing this backtracking process, we can recover P_i.

The algorithm given above has one significant inefficiency: in step 2, if all sums $s(p) + k(e)$ have values of at least $m' > m$, then the distance counter m should be increased immediately to m'.

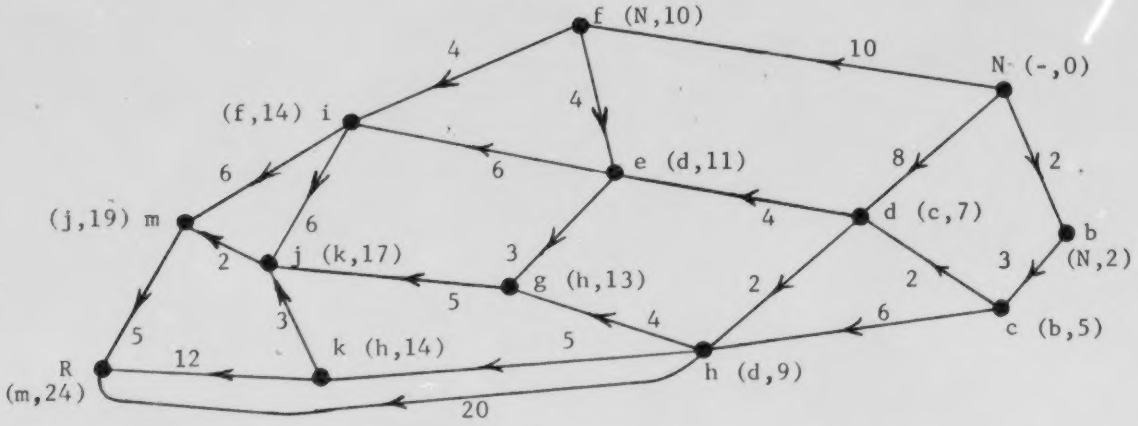


Figure 7

Example 1 - Shortest Path Problems:

A newly married couple, upon finding that they are incompatible, wants to find a shortest path from point N (Niagara Falls) to point R (Reno) in the road network shown in Figure 7. We apply the Shortest Path Algorithm. First N is labeled $(-,0)$. For $m = 1$, no new labeling can be done (we check edges (N,b), (N,d) and (N,f)). For $m = 2$, $s(N) + K(N,B) = 0 + 2 = 2$, and we label b with $(N,2)$. For $m = 3, 4$, no new labeling can be done. For $m = 5$, $s(b) + k(b,c) = 2 + 3 = 5$ and we label c with $(b,5)$. We continue to obtain the labeling shown in Figure 7. Backtracking from R, we find the shortest path to be $(N, b, c, d, h, k, j, m, R)$ with length 24.

If we want simultaneously to find shortest distances between all pairs of nodes (without directly finding all the associated shortest paths), we can use the simple algorithm due to Floyd. Let matrix D have entry $d_{ij} = \infty$ (or a very large number) if there is no edge from the i-th node to the j-th node; or else d_{ij} = the length of the edge from the i-th node to the j-th node. Then Floyd's algorithm is most easily stated by giving the FORTRAN code:

```
DO 1 K = 1, N
DO 1 I = 1, N
DO 1 J = 1, N
IF D(I,K) + D(K,J) < D(I,J) THEN
   D(I,J) = D(I,K) + D(K,J)
1 CONTINUE
```

When finished, d_{ij} will be the shortest distance from the i-th node to the j-th node.

b. Minimal Spanning Tree Algorithm.

We present two minimal spanning tree algorithms for connected, undirected n-node graphs. Note that any spanning tree of a connected, undirected n-node graph has $n-1$ edges (Exercise 30a).

Prim's Algorithm: Repeat the following step until the set S has $n-1$ edges (initially S is empty): add to S the shortest edge that does not form a circuit with edges already in S.

Kruskal's Algorithm: Repeat the following step until the tree T has $n-1$ edges: add to T the shortest edge between a node in T and a node not in T (initially pick any edge of shortest length).

In both algorithms, when there is a tie for the shortest edge to be added, any of the tied edges may be chosen. Showing Prim's algorithm does indeed form a spanning tree is left as an exercise. Note that Kruskal's algorithm is intuitively quicker because there are fewer edges to check in each iteration and no worry about forming circuits. Note also that Kruskal's algorithm is very similar to the Shortest Path Algorithm: if we consider nodes in T as labeled nodes, then both algorithms repeatedly search all edges from a labeled node to an unlabeled node (although the search is for different purposes). Indeed, the edges used to label new nodes in step 2 of the path algorithm form a directed spanning tree (Exercises 29).

The difficult part in the minimal spanning tree problem obviously is proving the minimality of the two algorithms. We give the proof for Kruskal's algorithm and leave Prim's as an exercise.

Theorem 3: Kruskal's algorithm yields a minimal spanning tree.

Proof: Suppose $T = \{e_1, e_2, \ldots, e_{n-1}\}$ is the spanning tree constructed by Kruskal's algorithm, with the edges indexed in order of their inclusion into T, and T' is a minimal spanning tree chosen to have as many edges in common with T as possible. We shall prove that $T = T'$. Assume $T \neq T'$ and let $e_k = (a,b)$, chosen on the k-th round of the algorithm, be the first edge of T (having smallest index) that is not in T'. Let $P = (e_1', e_2', \ldots, e_n')$ be the (unique) path in T' from a to b (in T, the path from a to b is simply e_k). If every edge on P is shorter than e_k, then on the k-th and later rounds, Kruskal's algorithm would have incorporated successive edges of P before considering the longer e_k. (A technical note: the algorithm could choose e_1' on the k-th round without fear of forming a circuit since edges chosen up to the k-th round plus the edge e_1' are all in the (circuit-free) tree T'.) If P has an edge with the same length as e_k, we remove this edge from T' and replace it by e_k. It is not hard to show that the new T' is still a spanning tree which has the same minimal length and which has one more edge in common with T - this contradicts the choice of the original T'. If P has an edge with greater length than e_k, we remove it from T' and replace it by e_k to get a shorted spanning tree--this contradicts the minimality of T'. Q.E.D.

Example 2 - Minimal Spanning Tree: We seek a minimal spanning tree for the network in Figure 8. Both algorithms start with a shortest edge. There are three edges of length 1: (a,f), (ℓ,q) and (r,w). Suppose we pick (a,f). If we follow Kruskal's algorithm, the next edge we would add is (a,b)

of length 2, then (f,g) of length 4, then (g,ℓ), then (ℓ,q), then (ℓ,m), etc. The next-to-last addition would be either (m,n) or (o,t), both of length 5 (suppose we choose (m,n)), and either one would be followed by (n,o). The choice of (m,n) or (o,t) brings out the fact that minimal spanning trees are not unique. The final tree is indicated with darkened lines. On the other hand, if we follow Prim's algorithm, we first include all three edges of length 1. Next we would add all the endges of length 2: (a,b), (e,j), (g,ℓ), (h,i), (ℓ,m), (p,u), (s,x), (x,y). Next we would add almost all the edges of length 3: (c,h), (d,e), (k,ℓ), (k,p), (q,v), (r,s), (v,w), but not (w,x) unless (r,s) were omitted (if both were present we would get a circuit containing these two edges together with edges of shorther length (r,w) and (s,x)). Next we would add all the edges of length 4 and finally either (m,n) or (o,t) to obtain the same minimal spanning tree(s) produced by Kruskal's algorithm. This similarity is no coincidence.

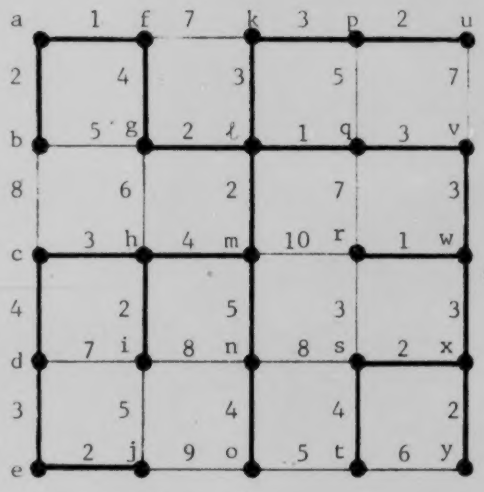


Figure 8

c. The Transportation Problem.

We are given an $n \times m$ matrix A of transportation costs, i.e., a_{ij} = cost of shipping one unit of our commodity from origin O_i to depot D_j, along with supply and demand vectors B and C, i.e., b_i = supply at O_i, c_j is demand at D_j. We assume $\sum b_i = \sum c_j$. The goal is to minimize the bill (total cost) for shipping the commodities to the depots. Remember that although the a_{ij}'s were lengths of time in the sweeping problem (and b_i and c_j were non-zero degrees of nodes), we can treat the minimal pairing of non-zero degree nodes as a transportation problem.

Let x_{ij} be the number of units sent from O_i to D_j. Then our problem is to minimize

$$\text{total costs} = \sum_{i,j} a_{ij} x_{ij} \; .$$

subject to the constraints

$$\text{(supplies)} \qquad \sum_{j=1}^{m} x_{ij} = b_1, \quad i = 1, 2, \ldots, n, \quad \text{and}$$

$$\text{(demands)} \qquad \sum_{i=1}^{n} x_{ij} = c_j, \quad j = 1, 2, -, m. \qquad (*)$$

In addition we want x_{ij} to be non-negative integers. It suffices to require that x_{ij} be non-negative, for all optimal solutions then they turn out to be integral (when B and C are integral). In this algebraic setting, the transportation problem has the form of what is called a linear program. While there are good algorithms for solving linear programs, we shall present an elegant algorithm due to Hitchcock that is much faster than general linear programming methods. See [4] for a more detailed discussion of this algorithm.

Our algorithm has 4 steps. First we seek an initial solution of positive x_{ij}'s that satisfy the constraints (*). We want to use as few positive x_{ij}'s as possible. We shall see that only $(n + m - 1)$ positive variables are needed. Our method for obtaining such a solution is best presented through an example. For the transportation problem given in Figure 9 (the costs are written in the top right box in each entry; the other numbers will be explained later), we choose x_{11} as large as possible. The first origin constraint implies $x_{11} \leq b_1$ and the first depot constraint implies $x_{11} \leq c_1$. So we set $x_{11} = \min(b_1, c_1)$. In this case $x_{11} = \min(40,30) = 30$. To try to use up the rest of the supplies of 0_1, we set $x_{12} = \min(c_2, b_1 - x_{11}) = \min(40,10) = 10$. Now we try to fill the rest of the demand at D_2 with x_{22}. We continue in this fashion setting $x_{22} = 30$, $x_{23} = 20$, and $x_{33} = 50$. Each new x_{ij} exhausts some supply or fills out some demand.

Apparently $n + m$ positive x's will be needed. However because $\Sigma b_i = \Sigma c_j$, when we use the last x_{ij} we must simultaneously satisfy both the last origin and last depot constraint. Thus $n + m - 1$ positive x's are needed. This technique for finding a solution to the constraints (*) is called the northwest-corner rule, because we zigzag our way across the matrix from the

		depot			
u_i		30	40	70	
origin	40	$\boxed{7}$ 30	$\boxed{3}$ 10	$\boxed{5}$ (0)	$u_1=0$
	50	$\boxed{6}$ (1)	$\boxed{1}$ 30	$\boxed{3}$ 20	$u_2=2$
	50	$\boxed{8}$ (0)	$\boxed{2}$ (-2)	$\boxed{6}$ 50	$u_3=-1$
		$v_1=7$	$v_2=3$	$v_3=5$	

Figure 9

northwest corner. See [4] for a further explanation of the northwest-corner rule and for modifications in the "degenerate" case where fewer than $n + m - 1$ positive x's are needed.

The next step involves writing down a solution to a "shadow" problem. In that problem, an outside shipper wants to determine two sets of prices; first, v_j, the price at which he will sell a unit of the commodity at D_j; and second, u_j, the price at which he will buy from us a unit at O_i. Then $v_j - u_i$ is the cost to us of shipping a unit from O_i to D_j. To do business with us, his transportation costs should not exceed the cost of our solution above. Since the shipper wants to make as much money as possible, he picks his costs to be exactly equal to ours for each x_{ij} used in the above solution. Thus his constraints are (i) $v_1 - u_1 = 7$; (ii) $v_2 - u_1 = 3$; (iii) $v_2 - u_2 = 1$; (iv) $v_3 - u_2 = 3$; and (v) $v_3 - u_3 = 6$. Since the prices are relative; i.e., only the differences are important, we can arbitrarily set $u_1 = 0$ (we always do this). Then (i) and (ii) determine $v_1 = 7$ and $v_2 = 3$. Then (iii) determines $u_2 = v_2 - 1 = 2$, and subsequently $v_3 = 3 + u_2 = 5$ and $u_3 = v_3 - 6 = -1$. Note that because we had only five equations (in general, $n + m - 1$ equations) in six unknowns (in general, $n + m$ unknowns), we had to determine one unknown arbitrarily. (In a "degenerate" case with less than $n + m - 1$ equations, more unknowns are set equal to 0; see [4].) With these prices, it now costs us as much to sell all our supplied to the shipper at the origins and buy the required amounts from him at the depots as it costs to use the solution obtained above.

The third step involves the choice of a new x_{ij} to be used to lower the cost of the existing solution. To beat the cost of our first solution satisfying the constraints (*), it suffices to get a solution that beats the cost of the outside shipper (since his total cost was our total cost). A natural first step is to compare our costs a_{ij} with the shipper's cost $v_j - u_i$ at each entry where x_{ij} is currently zero. In each such entry in Figure 9, we write $a_{ij} - (v_j - u_i)$ in parenthesis. We see that this difference is negative for engry (3,2). That means we save \$2 for each unit we ship internally from D_1 to D_2 instead of with the shipper (\$2 versus $3 - (-1) = \$4$). Then we alter the current solution so as to permit x_{32} to become positive. In general, we seek to increase the x_{ij} where $a_{ij} - (v_j - u_i)$ is most negative. If none of the differenes is negative, then there is no way to beat the shipper's cost (thus our own cost), and the current solution is minimal.

The fourth step determines the best improvement possible when we make the new x_{ij} positive. As we increase the new x_{ij}, in this case, x_{32}, we must balance the constraints (*) by changing other x_{ij}'s. We shall only be permitted to change the values of positive x_{ij}'s used in the current solution. The reason is that from step three we know that using other (currently zero) variables would either raise the cost of the new solution or not lower it as much. As we increase x_{32}, we must compensate by decreasing x_{22} and x_{33}. To compensate for the latter changes, we increase x_{23}. So $\Delta x_{22} = \Delta x_{33} = -\Delta x_{32}$ and $\Delta x_{23} = \Delta x_{32}$. Observe that a unit increase in x_{32} decreases the cost of the solution by $a_{22} + a_{33} - a_{32} - a_{23} = 1 + 6 - 2 - 3 = 2$ --this checks with the "predicted" savings in step three. The best improvement using x_{32} will obviously come from increasing x_{32} as much as possible. That is, we increase x_{32} until x_{33} or x_{22} becomes 0. So we can set $x_{32} = 30$ with $x_{33} = 20$, $x_{22} = 0$, and $x_{23} = 50$. The other x_{ij}'s are unchanged. Note that we apparently could have decreased x_{12} instead of x_{22} to balance the increase of x_{32}. Then we would need to increase x_{11} to balance the decrease in x_{12}, but now there is no way to balance this increase in x_{11}, since $x_{12} = x_{13} = 0$. It turns out that there is always one unique way to compensate for an increase in a previously unused variable. If the way is not obvious, we can find it by a systematic search. For further details, see [4].

Now we have the solution shown in Figure 10a. We repeat steps two, three, and four, and get the solution in Figure 10b. Now in step three, there is no entry that is less than the shipper's cost and so we have an optimal solution.

	30	40	70	
40	$\boxed{7}$ 30	$\boxed{3}$ 10	$\boxed{5}$ (-2)	$u_1=0$
50	$\boxed{6}$ (3)	$\boxed{1}$ (2)	$\boxed{3}$ 50	$u_2=4$
50	$\boxed{8}$ (2)	$\boxed{2}$ 30	$\boxed{6}$ 20	$u_3=1$
	$v_1=7$	$v_2=3$	$v_3=7$	

Figure 10a

	30	40	70	
40	$\boxed{7}$ 30	$\boxed{3}$ (2)	$\boxed{5}$ 10	$u_1=0$
50	$\boxed{6}$ (1)	$\boxed{1}$ (2)	$\boxed{3}$ 50	$u_2=2$
50	$\boxed{8}$ (0)	$\boxed{2}$ 40	$\boxed{6}$ 10	$u_3=-1$
	$v_1=7$	$v_2=1$	$v_3=5$	

Figure 10b

6. Summary of Procedure

The analysis developed in Section 4 can be divided into four general steps. First, we add a minimal-length set of edges so that the resulting graph G' has an Euler circuit in each of its components. Second, we pair inward and outward edges at each node so as to minimize unwanted turns and use these pairings to construct an Euler circuit for each component of G'. Third, if G' is not connected, then we link these circuits together in a grand tour. Fourth, we allow for various practical considerations as we manually break up the grand tour into routes of limited length. We assume the given graph G is contained in the larger graph H.

1. Appending edges to G to get a graph G' with Euler circuits in each of its components.
    - 1a. Obtain the matrix of approximate shortest distances in H between negative nodes x_i and positive nodes y_j of G. See page 116. (If relatively few nodes are involved, Minimal Path Algorithm is used to find shortest distances; see pages 123-125.
    - 1b. Solve the transportation problem with the matrix of step 1a and with row supplies $-d(x_i)$ and column demands $d(y_j)$. See pages 118 and 127-130.
    - 1c. Using the Minimal Path Algorithm, find the shortest paths in the larger graph H between the negative and positive nodes paired in step 1b. See pages 123-125.
    - 1d. Append dashed edges, duplicate and new, to G so that there are k dashed copies of an edge if that edge occurred on k of the paths in step 1c. Call the new graph G'. See pages 116-118.
2. Building Euler circuits in each component of G'.
    - 2a. Match up inward and outward edges at each node in G' by "inspection" or using the assignment problem approach. Use the weights in Figure 4. See pages 118-119.
    - 2b. Form the circuits arising from the match-ups in step 2a and paste these circuits together to get an Euler circuit in each component of G'. See pages 114 (Proof of Theorem 1) and 119.
3. Linking together the components of G' --performed only if G' is not connected.

138

    3a. Find the (approximate) shortest round-trip distances between the components of G'. See pages 119-120.
    3b. Using the distances in step 3a, find a minimal spanning tree for the component graph. See pages 125-127.
    3c. Find the shortest paths (in both directions) joining the closest pair of nodes in each pair of components linked in the minimal spanning tree. See pages 119-120.
    3d. Use these shortest paths to unite the tours of each component of G' to get the desired grand tour which covers all edges of G with minimal length.
4. Breaking up the grand tour.
    4a. Break up the grand tour into subtours of feasible length. At this point, the grand tour may be modified to allow for various constraints and other considerations. See pages 120-122.

7. An Example

Our procedure will now be illustrated with an example based on Figure 1 (page 111). The solid-lined edges in Figure 1b represent sides of streets on which parking is banned from 8:00 a.m. to 9:00 a.m. (the time period for our example). The graph G of the solid edges alone is given in Figure 11. The side of a street represented by an edge is indicated by the position of the edge relative to its end nodes in Figure 11. The dashed edges in Figure 1b represent the other streets in the district which may be used in deadheading. The time, in minutes, to sweep each edge is indicated in Figures 1b and 11. It is assumed that the time needed to deadhead an edge is one half of the time needed to sweep it.

Our analysis will procede as outlined in Section 6. Most of the computations are left as exercises and the results are given without explanation or are obtained by inspection and heuristics.

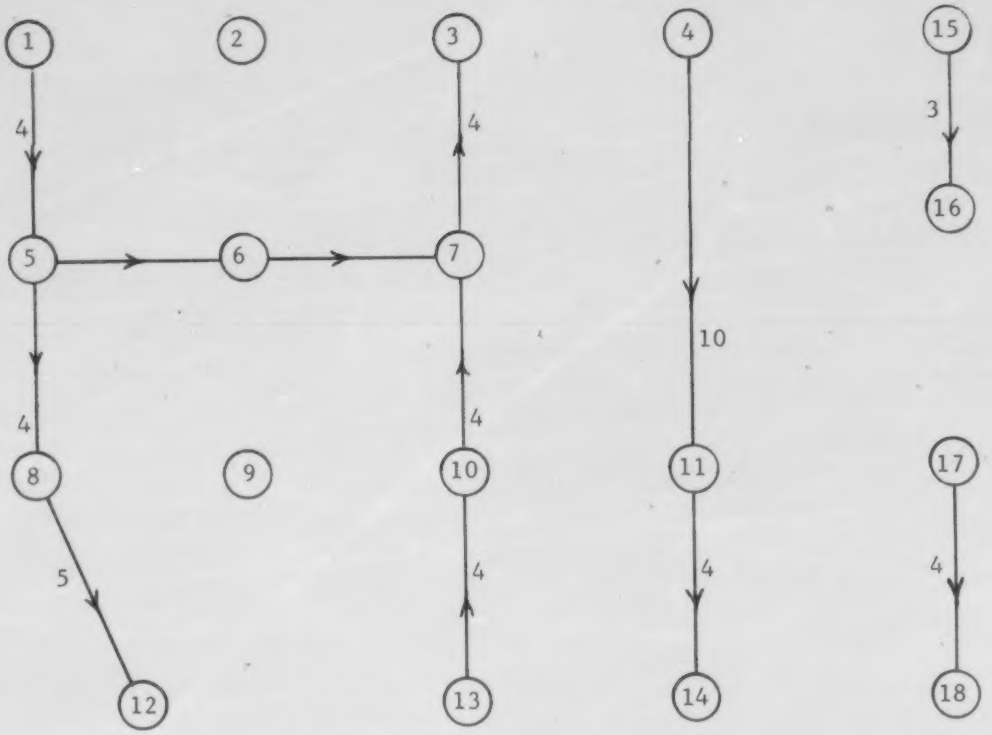


Figure 11

Step 1. We pick out the nodes of negative and positive degree which will be the origins and depots, respectively, of the transportation problem. The origins are nodes 3, 7, 12, 14, 16, 18. The depots are nodes 1, 4, 5, 13, 15, 17. The supply or demand at each of these nodes is one. (The unit supplies and demands will make the transportation problem "degenerate.") We must now compute the distance between each origin and depot. In the computerized analysis described in Section 4, the Manhattan distance was calculated from the nodes' coordinates. Because the graph in Figure 1b is relatively small, the coordinates of nodes have been omitted. The exact distances are readily obtained (if desired, coordinates can easily be invented). Most can be gotten by inspection and other distances can be ignored, as mentioned below. Figure 12 presents the matrix of relevant distances (students are asked to compute these distances in exercise 39). For simplicity, the distances are given in sweeping time, not deadheading time; times on deadheading edges will be divided by two later. A moment's thought makes it clear that nodes 16 and 15 and nodes

18 and 17 should be paired in an optimal solution to our transportation problem. Thus we can restrict the problem to the remaining four origins and four depots. The matrix in Figure 12 reflects this restriction.

		Depots			
		1	4	5	13
Origins	3	12	8	16	25
	7	16	12	20	22
	12	20	26	24	7
	14	32	28	36	8

Figure 12

One optimal solution to this transportation problem pairs node 3 with node 4, 7 with 1, 12 with 5, 14 with 13, as well as 16 with 15 and 18 with 17. (There exist other optimal solutions.) The sum of the lengths of the paths in this pairing is 63 minutes, or, at deadheading speed, $31\frac{1}{2}$ minutes. The deadheading edges, that is, the edges of the shortest paths between the paired nodes, are now added to the graph in Figure 11 to produce the graph G' in Figure 13. We easily check that every node in G' does have zero degree as desired. This completes step 1.

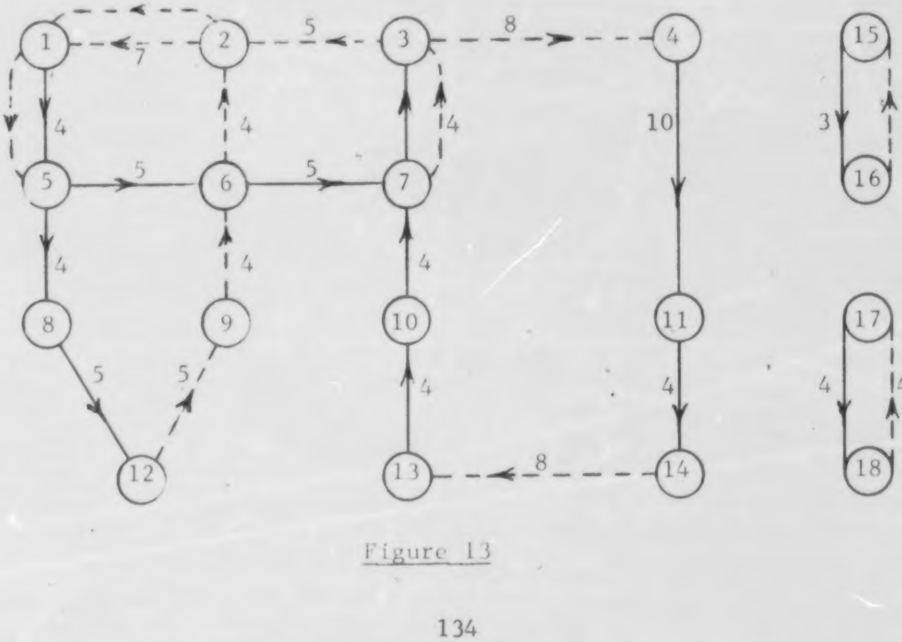


Figure 13

Let us note in passing some interesting properties of G': (i) G' has three components while G had four components; (ii) edge (2,1) is deadheaded twice; (iii) edges (1,5) and (7,3) are swept and deadheaded in the same direction.

Step 2. Each node in G' is now examined and the inward-outward edge assignment problem is solved. Only nodes 1, 2, 3, 5, 6, and 7 have two inward edges and at each of these an optimal assignment is obvious. The table in Figure 14a lists a set of optimal edge pairings at the above six nodes. Since only the nodes have names, the pairings are written as a triplet node sequence. We shall underline the parts of a sequence in Figure 14 that contain deadheading edges. From these pairings, we get a set of circuits described in the proof of Theorem 1. They are listed in Figure 14b. Note that, as is typical in most practical problems, there is little changing sides of a street. Moreover, there is no choice about how often or where to change sides--it must happen at nodes 4 and 14. Finally, we join the first and second circuits in Figure 14b; it does not matter whether we join them at node 7 or 3 for like circuits result. This completes step 2.

Node	Pairings
1	2-1-5, 2-1-5
2	6-2-1, 3-2-1
3	7-3-2, 7-3-4
5	1-5-8, 1-5-6
6	5-6-7, 9-6-2
7	6-7-3, 10-7-3

Figure 14a

Circuits

1-5-8-12-9-6-2-1-5-6-7-3-2-1

7-3-4-11-14-13-10-7

15-16-15

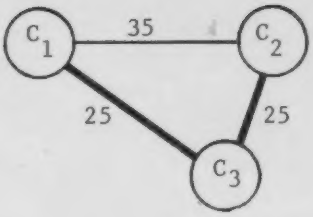
17-18-17

Combination of first two circuits

1-5-8-12-9-6-2-1-5-6-7-3-4-11-14-13-10-7-3-2-1

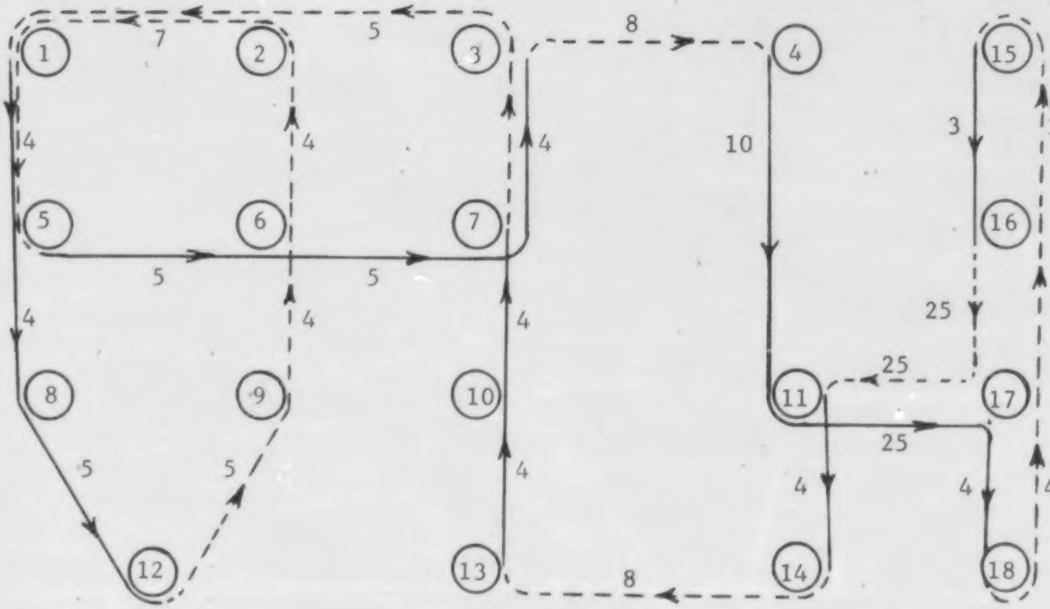
Figure 14b

Step 3. We need to find the shortest distances between the three components of G'. Let us call the large component C_1, the nodes 15-16 component C_2, and the nodes 17-18 component C_3. The closest nodes linking C_1 and C_2 are nodes 4 and 15 with distance 35. The closest nodes linking C_2 and C_3 are 16 and 17 with distance 25. The closest nodes linking C_1 and C_3 are 11 and 17 with distance 25. The minimal spanning tree among components is indicated with darkened edges in Figure 15. Using the corresponding edges, we connect the components to get the grand tour t: 1-5-8-12-9-6-2-1-5-6-7-3-4-11-17-18-17-16-15-16-17-11-14-13-10-7-3-2-1. The tour T is shown in Figure 16. It will take $137\frac{1}{2}$ minutes to complete this tour, 56 minutes of sweeping and $\frac{1}{2}(163) = 81\frac{1}{2}$ minutes of deadheading (remember that deadheading edges are traversed at twice the sweeping speed). This completes step 3.



C_1
C_2
C_3
35
25
25

Figure 15



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
7
5
8
4
4
4
10
3
3
5
5
4
4
4
25
25
25
25
5
5
4
4
4
4
8

Figure 16: Minimal tour covering all edges of graph in Figure 11.

Step 4. At first, it seem obvious that we cannot break the tour T into two 60-minute subtours, since T runs $137\frac{1}{2}$ minutes. There are many ways to break T into three feasible subtours. However it might be possible to break T so that long deadheading stretches around node 17 fell at the start or end of the subtours and thus could be dropped from the subtours. Clearly this is the only way we might be able to use only two subtours. Unfortunately, it is not quite possible to do this (the reader should verify this for himself). One way around this difficulty is to consider other ways to link together the components of G'. There are two other ways of linking components which while not giving a minimal-length grand tour do allow us to break the tour at long deadheading stretches so as to obtain two feasible subtours. If we combine components C_1 and C_3 of G' between nodes 14 and 18 instead of between 11 and 17, then the resulting grand tour would be T*: 1-5-8-12-9-6-2-1-5-6-7-3-4-11-14-18-17-16-15-16-17-18-14-13-10-7-3-2-1. This can be broken into subtours T_1: 1-5-8-12-9-6-2-1-5-6-7-3-4-11-14, and T_2: 15-16-17-18-14-13-10-7 of lengths 57 and $46\frac{1}{2}$ minutes, respectively. Note that by deadheading edge (1,5) the first time we traverse it (instead of the second time) in T_1, we get a further deadheading edge at the start of T_1. Thus we get T'_1: 5-8-12-9-6-1-5-6-7-3-4-11-14 of length 56 minutes. Together T'_1 and T_2 take $102\frac{1}{2}$ minutes with only $45\frac{1}{2}$ minutes of deadheading (T had $81\frac{1}{2}$ minutes of deadheading). If we link component C_1 of G' with component C_2 instead of C_3, we again get a grand tour containing two feasible subtours (the reader should find these feasible subtours). This concludes step 4.

8. Computer Implementation

While graphs are easy to represent (on paper) to the human eye, they are more difficult to represent for use inside a computer. As noted at the start of Section 2, we should represent edges as ordered pairs of nodes. So then the question is how do we represent nodes in a computer. The usual answer is as numbers. If a graph has n nodes, we use some scheme to assign each node a different number between 1 and n. Then we can read into the computer the set of edges--ordered pairs of nodes--for streets in a given district. There would also need to be secondary information about each edge: its length, side of street, and no-parking period. If we are going to estimate distances between nodes with the Manhattan distance, we need to read in the coordinates of each node. The coordinate information can also be used to check for errors in the node list as follows. Usually there exists a relatively small bound for the distance between any two nodes joined by an edge. We then program the computer to check that the distance between the ends of each edge is within this bound.

Inside the computer, we should reorganize the edges in terms of their end nodes. Thus for each node we want a list of incoming and outgoing edges. This information is just what we need in the shortest path and assignment problem calculations. The secondary information about each edge can be stored elsewhere. When a specific time period for sweeping is chosen, we can flag all the nodes and edges in the graph for that period. We still need to retain the other edges for possible use in deadheading.

Now students are ready to program the four steps in our sweeping procedure. A program for the whole procedure is a major undertaking and so some parts might be fudged. There are certain loose ends that a programmer must resolve. The main ones are:

1) How does one incorporate deadheading edges in a graph on a computer?
2) How does one store and represent circuits? How do you link them together?
3) Heuristics to find nearest pairs of nodes for pairs of components of G'.
4) Sometime you need to determine how many components the graph G' (with deadheading edges added) has. When? How? (Hint: the "when" and "how" are naturally related.)

Note that for step 4, the program should print out the grand tour for manual breaking up and also print out a minimal set(s) of feasible subtours found by the method presented on page

9. Summary and Extensions

In this module, we have presented a detailed mathematical procedure for routing mechanical brooms during a given period of time in a given district of a large city. At present, only a few cities in the country have the personnel and funds needed to utilize and maintain a computer program implementing this procedure. Only two cities, New York City and Washington, D. C., are now working with the procedure. Further, it takes a fairly large city for the procedure to improve over hand-drawn routes on the number of brooms needed in a given period in a given district. However, there are still several ways in which the model's analysis would be useful to a large city.

Even if the mathematically obtained set of subtours for a graph is not smaller than the present set, the individual tours would probably be shorter (with less deadheading time). Thus vehicles would be used less. The model's set of subtours serve as a good standard against which the manually generated routing can be rated. Most importantly, with such a model, one can quickly check out the results of various proposed changes in a city's parking regulations. (In New York City, almost every week sees some minor change in parking regulations in some of the districts.) The model can also be used to examine how changing the boundaries of the districts could result in a city-wide reduction in the number of brooms needed at peak demand. (We must note that the methodology for solving this city-wide problem is in a primitive stage; further, administrators at present feel that the complications involved in changing district lines would not be worth the possible savings.)

It should be recalled that there are many constraints that are difficult to quantify in urban routing problems. In the current problem we have incorporated all major constraints. Yet the precision of our model must not make us forget the uncertainties of day-to-day operations. For example, the time required to go down an edge (street) can only be an average time, varying with traffic congestion. The final step in which we manually check out the breaking of the grand tour into subtours permits us some practical adjustments to compensate for any over-precision of the mathematical analysis. We also should remember, as we saw in Section 7, that the set of subtours our procedure gives may not be minimal (but in practice so far, the computerized sets of tours have always been as good as hand drawn sets and usually have at least 20% less deadheading).

While we have included all major constraints of the original problem we were given, there are many ways to extend the analysis in this model. One could seek a more systematic method to paste together the circuits into an Euler circuit in each component of G', a method that minimizes unwanted turns. One could examine various ways of linking the components of G' in search of a grand tour in which feasible subtours can be made to start and end at the long deadheading stretches between components--as we did in the example in Section 7. One could consider the problem of linking up the routes of one period with the routes of the next period (perhaps, even design routes in one period to shorten this linking stage). Or one could incorporate into the model the time required to move from or to the district garage when dealing with sweeping periods at the start or end of a work period. If we permit a vehicle to travel along a curb in the opposite direction of the traffic flow when the vehicle is sweeping (few major cities permit this), then the edges of our graph become undirected (but deadheading would implicitly be directed). Most of our procedure would require substantial changes in this case.

In our model, we have pushed the mathematical analysis almost past the "state of the art" in urban science. Still, the resulting procedure has been used to achieve cost savings. We chose this model because it did show the state of the art and because it used a wide array of basic concepts and algorithms of operations research. Another model [8] developed at Stony Brook for New York City Department of Sanitation to optimize manpower scheduling contained less interesting mathematics but resulted in an annual savings of about $10,000,000 per year in New York City (the model even played a principal role in union contract negotiations).

10. Exercises

These exercises are divided into three sections. Exercises 1-10 present a variety of graph modeling. Exercises 11-38 deal with graph-theoretic concepts, variations and extensions of our analysis, and theory behind the algorithms. Exercises 38-53 contain numerical exercises.

1. In a football season, each pair of teams in a football league plays each other once. We assume that the result of each game is a win for one of the teams (no ties).

a) Describe how a directed graph can be used to represent the outcomes (who won) of each game in a season.

b) Suppose A, B, C, D are the teams in the league and that A beats B, D: B beats C, D; C beats A; and D beats C. Draw the associated graph.

c) A ranking of teams is a list of the teams in which the i-th team in the list beat the (i+1)-st team. Give a graph-theoretic interpretation in the associated graph of a ranking.

d) Using the graph-theoretic interpretation of a ranking, find a ranking for the league in b) by looking at the graph.

2. Consider a group of individuals who pass information (rumors) among themselves.

a) Describe how an undirected graph can be used to describe the "lines of communication" in this group, i.e., the various pairs of people who talk together.

b) Draw the graph for a group of five people in which each person talks with exactly two other people in the group.

3. Suppose that we need to match a set X of objects, say boys, with elements in another set Y of objects, say girls, and that each object x in X can only be paired with elements in the subset $N(x)$ in Y.

a) Describe how an undirected graph can be used to show which are possible pairs in a matching.

b) Suppose Bill likes Ann, Diana, and Lolita; Fred likes Nan, Carol and Lolita; John likes Carol and Lolita; and Harry likes Diana and Lolita. Draw the associated graph and find a matching that assigns each boy to a different girl.

c) A graph such as in a) is called bipartite--that is, the nodes divide into two parts and all edges go between, rather than within, the two parts. Show that an undirected graph is bipartite if and only if all its circuits are of even length.

4. A map of several countries is often colored so that countries with a common border (not just a common corner but a border of positive length) are drawn with different colors.

a) Describe how an undirected graph can be used to indicate which pairs of countries on the map have a common border.

b) Treat the network in Figure 7 as a map--the edges are the borders. Draw the associated graph of bordering countries.

c) Restate the condition for coloring countries of a map in terms of coloring nodes in the associated graph.

d) What is the minimum number of colors needed to color the nodes in the graph of bordering countries in b)?

e) Draw a graph that could not represent a set of bordering countries on map. Explain.

f) Draw two graphs such that each cannot be colored with just three colors and such that each could be 3-colored after any one node was removed.

5. A state legislature has many committees. Certain senior legislators are on several committees. Thus the memberships of the different committees overlap.

a) Describe how a graph can be used to represent which committees have overlapping membership.

b) Committee A has legislators 1, 3, 5, 6; B has 2, 4, 8, 10; C has 1, 7, 9; D has 2, 5, 8; E has 2, 4, 10, and F has 11, 12, 13. Draw the associated graph described in a).

6. A collection of garbage truck routes is drawn for a 2-day period so that each pick-up site is visited by one or two routes depending on whether the site needs daily or every-other-day service. A sample set of routes is shown in Figure 1E. We want to partition the routes between the first and second day of the period so that no daily site is visited both times on the same day.

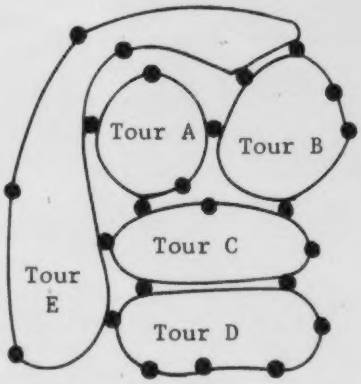


Figure 1E

a) Describe how a graph can be used to indicate whether or not any pair of routes can be assigned to the same day.

b) Draw the graph for the set of routes in Figure 1E.

c) What condition must this associated graph satisfy to guarantee the routes can be partitioned as required? Does the graph in b) satisfy this condition?

d) Restate the condition in c) in terms of a coloring of the nodes in the associated graph.

7. In a computer, a search to identify an unknown word, or for simplicity, an unknown letter, is performed as follows: The computer can test whether the letter is larger (in alphabetical order) than a given letter and one uses a scheme of such tests to identify the unknown letter. For example, if the unknown letter is one of A, E, M, X, then an efficient scheme would first test if the letter is greater than E and then one would ask if it is greater than A if the first answer were no, and if it is greater than M if the

first answer was yes. This scheme can be represented with a graph called a binary search tree with the branching process which narrows down the identification of the unknown letter. See example in Figure 2E.

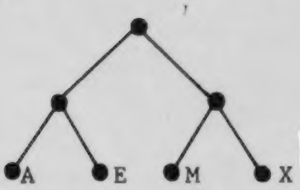


Figure 2E

a) Draw the graph associated with the testing scheme that asks is the unknown letter greater than A, then greater than E, then M, then X.

b) If each possible value of the unknown letter has a given probability of occurring, we can calculate the average number of tests needed to identify an unknown letter. If each of A, E, M, X has probability ¼, what is the average number of tests for a letter in the scheme (graph) in a)?

c) What graph minimizes the average number of tests when each letter has probability ¼? Prove your answer carefully.

d) Suppose the probabilities are: A 3/10, B 5/10, M 1/10, X 1/10. Now which graph minimizes the average number of tests?

8. Suppose we have an n x m rectangular chess board (instead of the standard 8 x 8 board) and we wonder whether there is a sequence of permissible steps by which a knight can go from one given square to another given square.

a) Describe how to draw an associated graph such that a path in the graph would represent a sequence of permissible knight moves.

b) Draw this graph for a 3 x 3 chess board.

c) Is there a sequence of permissible knight moves between every pair of squares on a 3 x 3 board? Explain in terms of the graph.

9. Some famous experiments about the structure of genes involved the following data. Many small segments along a certain gene had been identified and from the experiments one knew which segments overlapped. The general question was, when combined together did these segments form a long string (linear structure) or would the combined structure have circuits, branches, etc. In particular, one wanted to know if the overlap could possibly come from a linear structure or could it only be generated by a more complex structure.

a) Describe how to draw an associated graph that would represent the overlap information.

b) From the overlap information in Figure 3E (an "x" indicates overlap) draw the associated graph.

c) Draw some graphs which could not arise from overlapping segments on a line.

d) Could the graph in b) arise from overlapping segments on a line?

	1	2	3	4	5	6
1	x	x				x
2	x	x		x		
3			x	x	x	
4		x	x	x	x	x
5			x	x	x	
6	x			x		x

Figure 3E

10. In psychophysics, one attempts to define measures for such concepts as how appetizing various meals appear to a given person. Here one wants a preference function f that assigns to each possible meal x, a value f(x) such that the person prefers meal x to meal y if and only if $f(x) > f(y)+1$ The "+1" factor allows for the fact that it is hard for a person to discriminate between two meals that are different yet similar, e.g., half a pie versus 49/100th's of a pie. Conversely, a person would be indifferent between x and y if and only if $|f(x) - f(y)| \leq 1$. We can define an indifference graph to represent which meals are indistinguishable to the person.

a) How might a preference function be defined in terms of the indifference graph?

b) Suppose the matrix in Figure 3E tells which pairs of six meals are indifferent. Draw the indifference graph and find a preference function for that graph (if one exists).

c) Draw two different 4-node indifference graphs that have no preference functions.

11. Suppose we let nodes, instead of edges, represent sides of a street in a graph model for the sweeping problem.

a) Describe how the edges of this other graph should be defined. Where do the street 'lengths' go?

b) Draw this other graph for the road network in Figure 1a.

c) Instead of finding a minimal set of routes covering all edges, what do we want in this other graph?

12. Suppose instead of an Euler circuit, we wanted a Euler path--a path crossing every edge just once but not ending where we started. State and prove Theorem 1 for Euler paths in a directed graph.

13. Consider a connected network of two-way streets.

a) Prove that there exists a tour (a circuit) that traverses each edge once in each direction.

b) Pretend that all streets in Figure 1a are two-way. Now draw a tour of the type in a).

c) When tracing out a tour in b), let S_x denote the street just traversed as we arrive at corner x the first time. Prove that if one follows the rule: when at corner x, do not leave that corner along street S_x unless it is the only remaining possiblility--then you will always trace the tour in one step (no pasting circuits as in Theorem 1).

14. In an undirected graph, an Euler circuit is a circuit that traverses each edge once (in some direction, not both directions). State and prove Theorem 1 for undirected graphs.

15. What is the sum of the degrees of all the nodes in a directed graph? Prove your result rigorously.

16. If a directed graph has exactly two nodes of non-zero degree, prove that there is a path from one to the other.

17. If k is the sum of the degrees of all positive nodes in a directed connected graph G, prove that there exists a set of k paths such that each edge of G is on exactly one of these paths.

18. To get a tour covering all edges of a directed graph, we needed to add deadheading edges in order to make each node have zero degree.

a) What equivalent condition would deadheading edges have to satisfy if one were working with an undirected graph?

b) Restate and prove Theorem 2--for a minimal set of deadheading edges in an undirected graph which satisfy the condition in a).

19. Suppose that instead of a minimal-length circuit covering all edges, one wanted a minimal-length path covering all edges in a direct connected graph (assume the graph does not have an Euler circuit). Thus deadheading edges are added to get a graph with an Euler path. (See Exercise 12.)

a) Restate and prove the appropriate form of Theorem 2.

b) How should the transportation problem be modified to get the set of paths required by the revised form of Theorem 2 in a)?

c) Suppose that one wishes to find a minimal-length set of K (or fewer) paths covering all edges. Restate (but do not prove) the appropriate form of Theorem 2 and modify the transportation problem accordingly.

20. Prove that if an undirected graph with n nodes is a tree, then it has n - 1 edges.

21. Prove that if an undirected graph is a tree, then there is a unique path between any given pair of nodes.

22. An arborescence is the directed version of a tree--it is a tree when we ignore the direction of the edges and further all its edges are directed out from a given node called the root. See the example in Figure 4E. An inverted arborescence is like an arborescence except that all edges are now directed towards the root. A spanning (inverted) arborescence of a graph G is a subgraph that is an (inverted) arborescence and which contains all nodes of G.

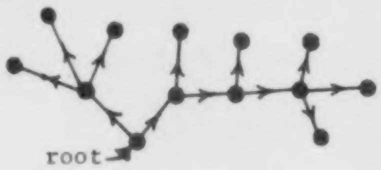


Figure 4E

a) Prove that if a directed graph satisfies the conditions of Theorem 1 then it has a spanning arborescence and inverted arborescence rooted at any given node. (Hint: use part of the Euler circuit.)

b) Prove that if there are paths in both directions between every pair of nodes in a directed graph, then there exists a spanning arborescence rooted at any given node.

23. Suppose that we have a directed graph satisfying the conditions of Theorem 1 and that we are given a spanning inverted arborescence A (see exercise 22) rooted at node x. Show that if we start tracing a path from x and only follow an edge of A when there is no other unused edge to leave the node we are currently at, then we will trace out an Euler circuit in one pass (with no pasting as in Theorem 1's proof).

24. Give a formal proof that the Shortest Path Algorithm on page 123. finds the shortest path from a to every other node in the graph.

25. Prove that Floyd's shortest path algorithm finds shortest paths.

26. a) Restate the Shortest Path algorithm on page 123 for

undirected graphs.

b) Ignoring the directions of the edges in Figure 7, find the shortest path from N to m.

27. Alter the Shortest Path Algorithm to find shortest path from each node to a node z in a directed graph.

28. Speed up the Shortest Path Algorithm by searching in step 2 among all edges from a labeled node p to an unlabeled node q for an edge that minimizes $d(p) + k(p,q)$. Explain how this would work.

29. Prove that in the Shortest Path Algorithm, the edges used in step 2 to label new nodes form a spanning arborescence (see exercise 22).

30. a) Prove that Prim's algorithm gives a minimal spanning tree.

b) Prove that the spanning tree of a connected, undirected n-node graph has $n-1$ edges.

31. Prove the following lemma used in the middle of the proof of Kruskal's algorithm: If T and T' are spanning trees, if $e_k = (a,b)$ is an edge in T but not in T', and if $P = (e'_1, e'_2, \ldots, e'_m)$ is a path in T' from a to b with e'_j being an edge on P not in T, then removing e_k from T and replacing it by e'_j will yield another spanning tree.

32. Modify Kruskal's algorithm so that it finds a minimal spanning tree that contains a prescribed edge e. Prove your modification works.

33. Modify Kruskal's algorithm so that it finds a maximal spanning tree.

34. Assume that if the edges of the undirected connected graph G are ordered, then when there is a tie in Kruskal's algorithm for the next edge to be added, the edge of smaller index is chosen. Recall that G often has several minimal spanning trees. Prove that the edges can be ordered so that Kruskal's algorithm, with the above tie-breaking rule, will yield any given spanning tree.

35. Describe how any transportation problem with integer supplies and demands can be converted into a transportation problem with supplies and demands all equal to one.

36. A transportation problem with unit supplies and demands can be converted into a matching problem in a bipartite graph (see exercise 3) in

which one seeks a matching which minimizes the sum of the lengths of the edges in the matching.

a) Describe this conversion.

b) Prove that the intermediate and final solutions obtained in our transportation problem algorithm all correspond to spanning trees in the associated bipartite graph of a).

37. Suppose that we have generated sets of routes in a given district for periods 1 and 2. There are four routes in each period and suppose x_i is the node where the i-th route in period 1 ends and y_j is the node where the j-th route in period 2 begins. One wants to pair up the routes of periods 1 and 2 so as to minimize the sum of the distances between the ends of the period 1 routes and the beginning of the period 2 routes. Pose this problem in a form which can be solved by one of the algorithms introduced in this module.

38. Suppose we are working in a district with no parking regulations (parking is never banned) and we can introduce any regulations we want, but parking can only be banned for one hour a day on any street. Describe how after solving the sweeping problem for a whole day (say, an 8 hour period), getting a collection of 8-hour routes that cover all edges, we can then set up "nice" (compatible) parking regulations.

39. a) Verify the distances in the matrix in Figure 12.

b) Use the transportation problem algorithm to solve the node pairing problem for nodes in Figure 12.

40. Find shortest paths in Figure 7 from f to R and from f to h.

41. Direct the edges in Figure 8 by the alphabetical order of the nodes, e.g., edge (h,m), not (m,h). Find the shortest paths from b to t and from a to y.

42. Find shortest paths in Figure 1b, using all edges:

a) from 1 to 14 b) from 14 to 1 c) from 5 to 15.

43. Program Floyd's algorithm for Figure 7 or manually use this algorithm for the subgraph in Figure 7 involving just nodes N, b, c, d, e, f, g.

44. Ignoring the directions of edges, find a minimal spanning tree for Figure 7. Is it unique?

45. Ignoring the directions of edges, find a minimal spanning tree for the whole graph (dashed and solid edges) in Figure 1b. Is it unique?

46. Find a spanning tree for Figure 8 which includes the edges (b,c) and (b,g) and which is of minimal length. (See exercise 32.)

47. Find a maximal-length spanning tree for Figure 8. (See exercise 33.)

48. Solve the following transportation problems:

a)

	60	30	30
40	6	3	4
40	3	1	7
40	5	3	7

b)

	50	30	90
70	5	2	6
40	3	1	8
60	7	4	9

c)

	10	3	2	3
5	6	4	3	2
9	5	2	4	4
4	4	2	2	4

49. In the end of the example in Section 7, it is claimed that a set of two feasible subtours can be obtained if one links component C_1 with C_2, instead of C_3, when building the grand tour. Find this set of two feasible subtours.

50. Consider the set of dashed edges in Figure 1b involving just nodes 1, 2, 3, 4, 5, 6, 7.

a) Generate a minimal-length circuit covering all these edges following the procedure outlined in Section 6.

51. Consider the set of dashed edges in Figure 1b involving just nodes 1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 13.

a) Repeat part a) in #50 for these edges.

b) Assuming a one-hour period, use your result in a) as a basis for finding a minimal set of feasible tours covering all these edges.

52. Consider the set of all dashed edges in Figure 1b.

a) Repeat part a) in #50 for these edges.

b) Assuming a two-hour period, use your result in a) as a basis for finding a minimal set of feasible tours covering all dashed edges.

53. Consider the graph in Figure 7. Suppose we add extra dashed edges (R,N) of length 25, (m,b) of length 20, and (f,N) of length 15.

a) Repeat part a) in #50 for the original solid edges in Figure 7.

b) Assuming a two-hour period, use your result in a) as a basis for finding a minimal set of feasible tours covering all solid edges.

REFERENCES

[1] Busacker, F. and Saaty, T. Finite Graphs and Their Applications. McGraw-Hill Book Company, New York, 1966--a good basic text in graph theory that covers a lot of ground; the second half is a survey of applications.

[2] Ore, O. Graphs and Their Uses. Random House, New York, 1963--a short, introduction to graph theory written for able high school students (and lower division undergraduates); of the many interesting applications, the logical puzzles (Sections 2.4 and 6.1) are especially enjoyable.

[3] Wilson, R. Introduction to Graph Theory. Academic Press, New York, 1972--a concise, well written undergraduate text.

[4] Hillier, F. and Lieberman, G. Introduction to Operations Research. Holden Day, San Francisco, 1967--the classic undergraduate operations research text; the algorithms used in this module are discussed in greater detail in Chapters 6 and 7.

[5] Wagner, H. Principles of Operations Research. Prentice-Hall, New York, 1969--the other standard operations research text; is a bit more advanced than [4]; algorithms used in this module are in Chapter 6.

[6] Bodin, L. "A Taxometric Structure for Vehicle Routing and Scheduling Problems," to appear in Journal of Computers and the Urban Society, 1974--this paper presents the cluster-first method and some node-covering problems (in this module we did edge covering) as well as a primitive version of the procedure presented in this module.

[7] Beltrami, E. and Bodin, L. "Networks and Vehicle Routing for Municipal Waste Collection," Networks 4 (1973), pp. 65-94--this paper describes two other garbage routing problems as well as a primitive version of the procedure presented in this module.

[8] Meckling, J. "Chart Day Problem: A Case Study in Successful Innovation," Journal of Urban Policy, Vol. 2, No. 2, November 1974--this paper, written by a New York City Sanitation Department administrator, describes some mathematical analysis which did not contain as interesting modeling as in this module but which resulted in an annual savings to the city of about $4,000,000.

[8] Edmonds, J. and Johnson, E. "Matching, Euler Tours, and the Chinese Postman," Mathematical Programming 5 (1973), pp. 88-124--this paper presents a way to build an Euler circuit in a directed graph all at once (without first getting several circuits that must be joined together as we did in Theorem 1); their method is based on a directed spanning tree and is quite intuitive.

Chapter 5

A MATHEMATICAL MODEL OF RENEWABLE RESOURCE CONSERVATION

Colin W. Clark
The University of British Columbia

INTRODUCTION

Most people exhibit at least a partial interest in the question of conservation of resources. "Conservation" is one of those catch-all words--like "democracy," "liberation," etc.--which, while it denotes something that everyone considers desirable, can be used to mean several different things.

For members of the Sierra Club, or the Audubon Society, conservation is probably synonymous with preservation. In this sense, forest conservation would not be compatible with logging. But to a professional forester, conservation would probably denote some concept of optimal forest production. Logging would be the purpose of conservation.

In these notes we shall use the word "conservation" primarily in the latter sense. More precisely, let us adopt the (tentative) definition that conservation means the optimal use of resources. Actually, this definition should also satisfy Sierra Club members, who will no doubt prepare arguments to show that for many forests, the optimal use is a recreational one requiring preservation in the natural state.

The model to be discussed in these notes, however, is based on the opposite premise: a certain biological resource is to be harvested for the commercial market. What is the optimal harvest policy? This still leaves us with a difficult definition problem--what do we mean by "optimal?"

We shall adopt a completely businesslike interpretation: optimal resource exploitation means the maximization of economic revenues obtained therefrom, in a sense to be described more fully in Section 1. Our purpose is not to "sell" this admittedly one-sided definition, but rather to study some of its implications. Indeed, we shall conclude eventually that the definition may have severe limitations as an acceptable criterion for resource conservation.

Now let us consider the conservation question (in our agreed formulation) a little more seriously. A significant feature of resource exploitation is the effect that current use has on future supplies. An exhaustible resource, for example, can be used up "too quickly." A renewable resource can be "overexploited," in the sense that its productivity falls below its maximum level. Thus any problem of optimal resource management is basically a dynamic optimization problem. Such problems are theoretically among the most challenging and difficult faced by applied mathematicians. An entire field of mathematical analysis is devoted to them; it goes under the name of the "calculus of variations", or in its contemporary form, "optimal control theory."

Copyright © 1974 by Colin W. Clark. Reproduced with permission.

Don't despair! The reader of these notes is not assumed to know anything at all about optimal control theory! On the contrary, the only prerequisite to studying the notes is a good background in elementary calculus (including continuity arguments, and the like). Hopefully the notes will provide some readers with the motivation to study optimal control theory (and its cousin, dynamic programming) more seriously. No mathematician interested in today's real-world problems can afford to remain ignorant of this exciting field, which, by the way, is one of the most active areas of contemporary research.

The bio-economic model to be described in the following pages incorporates a large number of simplifications, many of which may not be justified in practical problems. It is of course the very nature of scientific study to begin with simplified models, adding more realistic components and increasing the complexity of the models as the subject is developed.

In the limited confines of these notes we can only follow this scientific process through one or two steps, which will nevertheless carry us far beyond the traditional models of renewable resource harvesting. Further ramifications of the subject will be found in my forthcoming book, "Mathematical Bioeconomics", to be published by Wiley-Interscience, and in the references cited at the end of the notes.

A word about the Exercises listed at the end of each section. Most of these are quite simple, but illustrate various points not discussed in the text. If you don't have time to work all the exercises, I hope you will at least read them through, and solve those that interest you most.

1. ELEMENTARY PRINCIPLES OF COST-BENEFIT ANALYSIS

A business executive wishes to know whether his firm should undertake a significant expansion; a farmer considers devoting 40 acres of his farm, now a hay field, to an apple orchard; an investor wonders whether $62.50 is a reasonable price to pay for a $100 municipal bond that will mature in 10 years' time. These problems all have a similar structure: a current expenditure, or "investment," is expected to result in certain monetary benefits in future years. Will the benefits be sufficiently large to make the investment "profitable?"

Problems of this kind can be dealt with by the techniques of cost-benefit analysis, which have been developed by economists to handle both private and government investment decision problems. The fact that the expected benefits and the expected costs may be extremely difficult to estimate accurately does not lessen the importance of performing such analyses. A firm--or a government--that habitually makes losing investments can only expect to become progressively poorer as the result.

In most investment problems, the expected costs and benefits will be spread out over a time span of significant duration. The three examples mentioned above are quite typical in having an initial cost (often thought of as "the investment"), followed by a stream of benefits. This characteristic brings in the fundamental notion of time discounting. In comparing the future benefits with present costs, the benefits will be discounted according to some selected discount rate, thereby reducing them to their present value. Let us define these terms more precisely.

The present value of a payment P due to be made n years from now, at an annual discount rate i, is by definition equal to the amount Q that must be invested now at compound interest with annual interest rate* i, in order to attain the value P in n years' time. Thus

$$P = Q(1 + i)^n$$

or

$$Q = \text{Present Value of } P = \frac{P}{(1 + i)^n} . \qquad (1.1)$$

It is clear that the present value of a given payment P decreases with the discount rate, i, and also with the time to maturity, n. Since investment decisions are invariably made in terms of discounted values, it follows that such decisions will also depend--often quite critically--on the interest rate i and on the length of the investment period n.

The purchase of a bond is a prototype of the simplest problem of this kind. The present value of a \$100 bond due in 10 years' time is

$$Q = \frac{\$100}{(1 + i)^{10}} .$$

This expression is tabulated below as a function of i:

Interest rate (i)	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
Present value (Q)	\$90.53	\$82.03	\$74.41	\$67.56	\$61.39	\$55.84	\$50.83	\$46.32	\$42.24	\$38.55

---

* Note that the terms "discount rate" and "interest rate" are synonymous.

The quoted price of $62.50 thus corresponds to a rate of interest between 4% and 5% (in fact, 4.81%). The investor will compare this rate of return with that available from alternative investment opportunities. Although other considerations, such as the security of the investment, may also affect his decision, the investor will tend to favor purchase of the bond whenever that appears to be his most profitable investment opportunity.

In making investment decisions, then, the investor will normally have in mind a given rate of interest determined by his alternatives. Economists refer to this rate as the "opportunity cost of capital," since it is determined by one's investment opportunities. The relationship, if any, between the opportunity-cost discount rate and a socially optimal discount rate is currently a subject of hot debate in economic circles. References to the literature are given at the end of these notes.

Instead of a single payment P in year n, most business investments yield a sequence of payments $\{P_k\}$ occuring in years $k = 1, 2, 3, \ldots$. The present value of such a sequence is then the sum of the present values of the individual payments:

$$\text{P.V.} = \sum_{\text{all } k} \frac{P_k}{(1+i)^k} . \qquad (1.2)$$

If $C_o = -P_o$ denotes the cost of the initial investment, and if generally $P_k = R_k - C_k$ denotes the net revenue in year k, then

$$\text{N.P.V.} = \sum_{k=0}^{N} \frac{P_k}{(1+i)^k}$$

represents the net present value of the investment opportunity, with respect to the time-horizon N and discount rate i.

The basic decision rule of cost-benefit analysis may be stated: invest if $\text{N.P.V.} > 0$. More generally when several opportunities are under consideration, the rule becomes: maximize N.P.V. This is the rule we shall apply in the next section to a resource-management model.

Exercises

(1.1) The farmer mentioned in tne first paragraph estimates costs and benefits of planting his 40-acre hay field in apple trees as follows.

Cost of planting in year 0	$5000
Cost of pruning, etc., in years 1, 2, ..., 40	$100 per year
Revenue from apple sales, year 11-40	$1000 per year

On the other hand, the farmer values the hay production from the field at $100 per year. If $i = .05$, which use is superior? What if $i = .10$? Use a 40-year time horizon.

(1.2) The instantaneous (or continuous) interest rate δ corresponding to an annual rate i is defined by

$$e^{\delta t} = (1 + i)^t$$

so that $\delta = \ln(1 + i)$.

Suppose there is a constant instantaneous rate of inflation δ_{inf}. Let δ_{obs} denote the observed instantaneous rate of interest. Explain why the "real" instantaneous rate of interest δ_{real} is given by

$$\delta_{real} = \delta_{inf} - \delta_{obs}.$$

(It is imporant to realize that it is the real rate of interest that should normally be utilized in cost-benefit analysis.)

(1.3) How long should wine be aged? Suppose that a wine merchant has in his cellar a certain stock of fine wine, the market value of which is given by a known function $V(t)$, where $t \geq 0$ denotes the age of the wine. Show that the "optimal" age t^* for selling off the wine satisfies

$$\frac{V'(t^*)}{V(t^*)} = \delta.$$

Show also that, if the merchant disposes of the wine at age t^* and deposits the proceeds in an account at (instantaneous) interest δ, he will maximize the size of the account at any given future time T.

2. A MODEL OF RESOURCE HARVESTING

We wish now to apply the principles of cost-benefit analysis to the question of renewable resource harvesting. Because our approach is "strategic" (i.e., theoretical) rather than "tactical" (i.e., practical), we shall utilize a simple generalized model of biological population dynamics. The analytical study of more complex and realistic biological models is difficult. Practical problems are frequently analyzed numerically, using the method of "dynamic programming," which is the general technique that underlies the material in this section. This method will be outlined briefly in Exercises (2.3) and (2.4).

Let x_k denote the size of a particular population of animals (fish, birds, game animals, etc.) in year k. The variable x_k may be measured in purely numerical units (the total number of animals in the population), or more commonly in units of mass. In the latter case x_k is referred to as the biomass of the population. Since we shall not consider any structural characteristics of the population, such as age, weight or sex of the individual animals, it does not matter here what units x_k represents.

We suppose that the size of the population in year $k+1$ is determined as a known function of the size in year k:

$$x_{k+1} = F(x_k) \tag{2.1}$$

where $F(x)$ is a function assumed (for simplicity) to satisfy the following hypotheses:

$$\left.\begin{array}{r} F'(x) > 0, \quad F''(x) < 0 \quad \text{for all} \quad x > 0; \\ F(0) = 0. \end{array}\right\} \tag{2.2}$$

In order to have a viable population model, we also require that

$$F'(0) > 1 \quad \text{and} \quad \lim_{x\to\infty} F'(x) < 1. \tag{2.3}$$

Thus the curve $y = F(x)$, which is called the growth curve of the population, has the appearance shown in Figure 1. There is a unique equilibrium

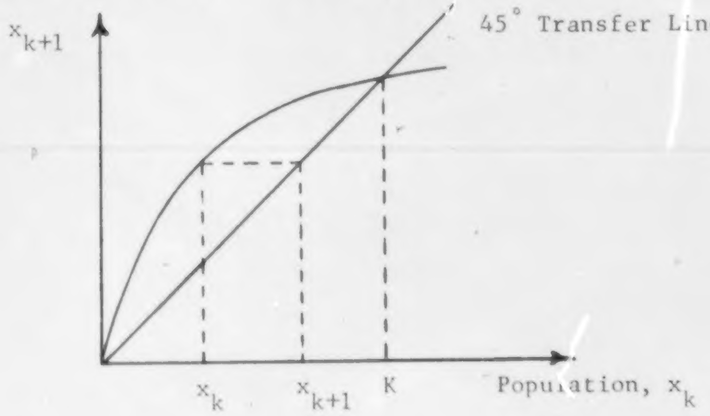


Figure 1

population $K > 0$ such that

$$F(K) = K.$$

The behavior of a sequence $\{x_k\}$ of population levels determined by Equation (2.1) and a given initial population x_1, is easily deduced from Figure 1. The 45° "transfer line" is used to transfer $x_{k+1} = F(x_k)$ back to the horizontal population axis. It is clear that the sequence $\{x_k\}$ approaches the equilibrium population K monotonically (because $F(x)$ is assumed increasing: cf. Exercise 2.1). Thus the equilibrium at $x = K$ is stable. There is another equilibrium at $x = 0$; mathematically this is an unstable equilibrium, although in a biological sense extinction may be very stable indeed!

We next introduce harvesting into our model. Suppose that a harvest h_k in year k reduces the "parent" population x_k to $x_k - h_k$, and that the reduced population is the breeding stock (or "escapement") from which the subsequent parent stock is derived:

$$x_{k+1} = F(x_k - h_k). \tag{2.3}$$

The harvests h_k are obviously constrained by the inequalities

$$0 \leqq h_k \leqq x_k, \quad k = 1, 2, 3, \ldots \tag{2.4}$$

A harvest sequence $\{h_k\}$ satisfying these inequalities is called a feasible harvest sequence.

The foregoing model describes quite realistically what takes place, for example, in the Pacific salmon fishery. The adult salmon spawn in costal streams, and die upon spawning. The young salmon spend a year or more in fresh water, eventually heading for the open ocean. After a period that varies from two to four years, depending on the species, the salmon return to the river of their birth in order to spawn. It is during the approach to this spawning run that the salmon are harvested along the coastline.

With $\Delta k = 1$ representing the length of the 2-4 year "cycle," our model is a good description of this fishery. The function $F(x)$, in this case called the "stock-recruitment" relationship, has been studied extensively by fisheries biologists.

Given the initial population x_1 and a specific sequence of feasible harvests $\{h_k\}$, the recursion formula (2.3) determines all future population levels x_k.* By a management policy we shall mean the choice of a feasible harvest sequence. In most instances, the choice will be determined so as to achieve some specified objective.

We are particularly interested in objectives suggested by the principles of cost-benefit analysis--that is, the maximization of the present value of net revenues. To what extent such an objective may represent a socially optimal harvesting policy, is a question that will be deferred until later.

Maximum Sustainable Yield

Before discussing the maximum present value (MPV) objective, however, we should pause to describe the more traditional objective of maximum sustainable yield (MSY). Let

$$G(x) = F(x) - x . \qquad (2.5)$$

Note that (Figure 2) $G(x)$ represents the sustainable yield corresponding to

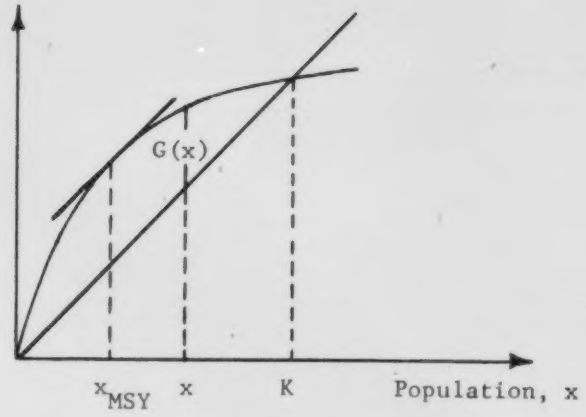


Figure 2

a given escapement population level equal to x. Maximum sustainable yield occurs at the point x_{MSY} where $G'(x) = 0$, i.e.

$$F'(x_{MSY}) = 1. \qquad (2.6)$$

---

* The student should observe the similarity between such first-order recursion schemes and first-order differential equations $dx/dt = f(x)$.

It is customary to refer to MSY as the "biologically optimal" level of exploitation, although the reasons behind this definition are rather obscure. Similarly, we may consider the population as being biologically overexploited whenever the escapement level x is less than x_{MSY}.

MSY is the explicitly recognized management objective for many national and international resource-regulating agencies. It is a simple objective that is readily understood; it obviously implies resource conservation by its very definition. Economists, however, have observed that there is no reason to suppose that MSY would be economically desirable. Moreover, observation of existing renewable resource industries, even where they may be under private ownership, shows that MSY is seldom encountered in practice, except in some cases under strict government regulation.

Maximum Present Value

Private resource owners, we may suppose, would tend to choose maximum present value (MPV) as their management objective. What effect would this have?

Let us assume for the moment that the net revenue from harvesting is proportional to the size of the harvest:

$$\text{Net Revenue} = ph_k$$

where p = price = constant. (This oversimplistic assumption will be relaxed in Section 3.) Then with a finite time horizon of N years, and a harvest sequence $h_1, h_2, \ldots, h_N$, the present value of net revenues equals

$$\text{P.V.} = p \sum_{k=1}^{N} \frac{h_k}{(1+i)^{k-1}} . \tag{2.7}$$

(We are now supposing that the first harvest $(k = 1)$ is taken immediately, so that the discount factor is $(1+i)^{k-1}$ rather than $(1+i)^k$. This simplifies the notation slightly. We also neglect any initial investment in harvesting facilities.)

We now have a somewhat nontrivial mathematical problem: determine a sequence $h_1, \ldots, h_N$ subject to conditions (2.3) and (2.4), so that the expression (2.7) attains a maximum value. The reader, if he has time, may enjoy tackling this problem on his own--give yourself an hour, say; the effort may well pay off by making it easier to follow the text.

Perhaps you began (as I did) with the trivial case $N = 1$: maximize ph_1 for $0 \leq h_1 \leq x_1$. The solution, obviously, is: $h_1^* = x_1$. When the time horizon is one period (next year is the end of the world!), it is optimal to harvest the whole population. This is obviously the correct solution, given our assumptions--but it may also suggest several glaring weaknesses in the assumptions. Some of these weaknesses are taken up later.

Next let $N = 2$; the problem now is*

$$\max\{h_1 + \alpha h_2\}, \quad \alpha = \frac{1}{1+i} \tag{2.8}$$

subject to conditions

$$0 \leq h_1 \leq x_1$$

$$0 \leq h_2 \leq x_2 = F(x_1 - h_1).$$

Suppose temporarily that h_1 is given; what is the optimal choice for h_2? When the harvest h_2 occurs there is only a one-year time horizon. We know the answer to this problem: $h_2^* = x_2 = F(x_1 - h_1)$. Therefore we can reduce our problem to:

$$\underset{0 \leq h_1 \leq x_1}{\text{maximize}} \{h_1 + \alpha F(x_1 - h_1)\}. \tag{2.9}$$

This is a simple one-variable maximization problem. Write

$$y = x_1 - h_1$$

so that (2.9) is equivalent to

$$\max_{0 \leq y \leq x_1} \{\alpha F(y) - y\}.$$

By calculus, the solution $y = q_i$ satisfies

$$F'(q_i) = \frac{1}{\alpha} = 1 + i \tag{2.10}$$

provided this equation has a solution q_i between 0 and x_1. The optimal harvest is then

$$h_1^* = x_1 - q_i.$$

In case $q_i > x_1$ it is easy to see that

$$h_1^* = 0$$

---

* We now set $p = 1$, since the value of p obviously does not affect the maximization problem.

is optimal. Thus in general we have

$$h_1^* = \max (0, x_1 - q_i). \tag{2.11}$$

Equation (2.10) fails to possess a solution $q_i \geqq 0$ if and only if $F'(0) < 1 + i$. In this case the maximum is at $y = 0$, i.e. $h_1^* = x_1$. This solution is included in formula (2.11) if we extend the definition of q_i by setting

$$q_i = 0 \quad \text{if} \quad F'(0) < 1 + i. \tag{2.12}$$

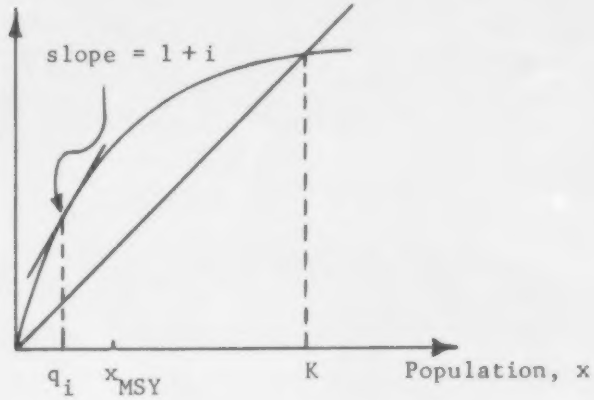


Figure 3

To summarize, the optimal escapement population (for a two-year time-horizon problem) is given by $x = q_i$, where q_i is defined by (2.10) or (2.12). The optimal first-year harvest policy reduces the initial population x_1 to q_i, if $x_1 > q_i$; otherwise the optimal harvest is zero (Figure 3).

It may not be obvious yet, but this first-year policy turns out to be optimal for any time horizon $N \geqq 2$. Thus, assuming $x_1 \geqq q_i$ we will have $h_1^* = x_1 - q_i$, and therefore

$$x_2 = F(x_1 - h_1^*) = F(q_i).$$

Hence, similarly

$$h_2^* = x_2 - q_i = F(q_i) - q_i$$

(except when $N = 2$, where we have $h_2^* = x_2 = F(q_i)$), so that

$$x_3 = F(x_2 - h_2^*) = F(q_i),$$

and so on. Consequently $x = q_i$ is the optimal equilibrium escapement population for every year except the last, and

$$h_i = F(q_i) - q_i = G(q_i)$$

is the optimal sustained yield for every year except the first and the last. (If $x_1 < q_i$, however, this sustained yield may not occur for several years, since $h_k^* = 0$ whenever $x_k < q_i$.) The correctness of this solution to our present-value maximization problem will be proved rigorously in Section 3. Further details appear in the Exercises at the end of the present section.

What is the economic significance of the rule (2.10), which determines the optimal escapement? This rule can be written in the form

$$G'(q_i) = i \tag{2.13}$$

where $G(x) = F(x) - x$ is the sustainable-yield function introduced earlier. We may also consider $G(x)$ as the (net) productivity of our population, as a function of the escapement level x. In economic terminology, $G'(x)$ is then called the "marginal" productivity of the population. Equation (2.13) then asserts that the optimal escapement level $x = q_i$ is determined by the condition

marginal productivity of resources = rate of interest.

This rule, which appears in many forms in economic analysis (cf. Exercise (1.3)), has a convincing intuitive explanation. Suppose harvesting has reduced the population level to x; the question arises whether to harvest an additional unit $\Delta x = 1$ (called the "marginal" unit by economists). Harvesting $\Delta x = 1$ will provide an immediate revenue of 1 unit, since we have assumed that $p = 1$. This unit of revenue can be invested at the given rate of interest i, to produce an annual income of i units.

At the same time, harvesting $\Delta x = 1$ will reduce the sustainable yield by an amount

$$G(x + 1) - G(x) \approx G'(x).$$

Thus the marginal annual benefit from harvesting equals i, and the marginal annual loss equals $G'(x)$. If $x > q_i$ then $G'(x) < i$ and the harvesting benefit is greater than the loss: further harvesting is profitable. Conversely, if $x \leqq q_i$, further harvesting is not profitable. Consequently $x = q_i$ is the optimal population at which to cease harvesting.

We have now reached the rather surprising conclusion that optimal resource harvesting (where "optimal" is understood in the normal MPV sense) is determined by the rate of interest--and nothing else! The "nothing else," of course, is a result of the model, which indeed contains no other economic parameters that could affect the solution. Another important parameter, cost, will be considered in the next section.

Dependence upon the interest rate

How does the solution q_i, and more significantly, the optimal sustained yield $G(q_i)$, depend on the interest rate i? The reader can easily verify, from the given assumptions, that:

(i) For $i = 0$ we have $q_o = x_{MSY}$, and thus $h_o = G(q_o)$ is the maximum sustainable yield.

(ii) As i increases, both q_i and h_i decrease monotonically, until

(iii) When $i \geq i_{crit} = G'(0)$, both q_i and h_i become zero.

The last conclusion asserts that extinction--in fact immediate extinction--is "optimal" whenever the maximum productivity of the biological population is less than the given rate of interest. From this point of view, biological resources that cannot produce at a sufficient rate, in comparison with the return yielded by other "investments," are simply expendable. The great whale populations are examples of resources that may be expendable in this sense (see Section 4), and so perhaps are many forests (although our model does not cover the latter case).

The reader should realize however, that the above analysis is primarily descriptive, rather than normative--even though the word "optimal" may seem to carry normative content. We have been asking the question: how would the private resource owner manage the resource, assuming the usual profit-maximization motive? We have tacitly assumed that the owner has no separate "preservation" motive that would inhibit him from destroying his resource stock, if such action turned out to be sufficiently profitable. Observation of many existing resource industries, such as whaling and other fisheries, lumbering, and cattle grazing, suggests that this assumption may not be too unreasonable. Many renewable resources tend to suffer from "biological overexploitation," in the sense that they eventually reach a state at which productivity is far below MSY. In extreme cases, extinction could be the final result.

Our results show, therefore, that serious biological overexploitation might be "explained" by means of high discount rates. An important alternative explanation will be described in the following section.

As mentioned earlier, economists have long argued whether the interest rate as determined by the investment market, i.e., the opportunity-cost rate, has any social significance. It has been proposed, for example, that since society has an obligation to future generations above and beyond that which private individuals are likely to observe, it follows that the *social* interest rate (often called the *social time-preference rate*) will be smaller than the private rate. Private industry will then be expected to practice less "conservation" than is socially desirable.

It is probably a mistake, however, to carry this argument to an extreme, insisting that society should adopt a zero time-preference rate. Russian economists, while prohibited by Marxist doctrine from refering to the capitalistic term "discount rate," have found it necessary nevertheless to utilize some form of the concept in their planning.

Although it may appear that a zero discount rate would be equitable, simply equating the present with the future, it is not hard to see that in fact the effect of a zero rate would be a total disregard for people's present welfare. For when $i = 0$, any present sacrifice that results in a future gain, *no matter how remote*, is desirable. Most people would agree, I think, that this is carrying intergenerational philanthrophy too far.

Exercises

(2.1) Drop the assumption that $F(x)$ is increasing, but retain the concavity assumption $F''(x) < 0$, and also assume $F'(0) > 1$. Then there is a unique equilibrium $K > 0$ for which $F(K) = K$. Show that K is a *stable* equilibrium if $F'(K) > -1$, but not if $F'(K) < -1$. {K is a stable equilibrium if $\lim x_k = K$ whenever x_1 is sufficiently close to K.} [HINT. Use the mean value theorem to show that $\{x_k\}$ is a Cauchy sequence if $F'(K) > -1$ and x_1 is near K.]

(2.2) Drop the assumption that $F(x)$ is concave, but assume it is increasing. Suppose there are two equilibria $0 < K_1 < K_2$ such that $F(x) < x$ for $0 < x < K_1$ and for $x > K_2$, and $F(x) > x$ for $K_1 < x < K_2$. Show that K_1 is then an unstable equilibrium, and 0 a stable equilibrium. The population level $x = K_1$ is sometimes called a *minimum viable population*.

Exercises (2.3) and (2.4) show how the solution for the case $N = 2$ given in the notes can be extended to arbitrary N. These exercises provide an introduction to the important technique of dynamic programming.

(2.3) Let $P_N(x_1)$ denote the maximum present value for time horizon N and initial population x_1:

$$P_N(x_1) = \max \sum_{k=1}^{N} h_k \alpha^{k-1}$$

where $\{x_k\}$ and $\{h_k\}$ satisfy (2.3) and (2.4). Show that

$$P_2(x_1) = \begin{cases} x_1 - q_i + \alpha F(q_i) & \text{if } x_1 \geqq q_i \\ \alpha F(x_1) & \text{if } x_1 < q_i \end{cases}$$

Thus show that $P_2(x_1)$ is a continuous increasing function of x_1. Finally show that $P_2(x_1)$ is a nonincreasing function of i. [Consider the graph.] It is usually true that the value of an asset decreases as the interest rate increases--recall the $100 bond of Section 1.

(2.4) Show that

$$P_{N+1}(x_1) = \max_{0 \leqq h_1 \leqq x_1} \{h_1 + \alpha P_N(F(x_1 - h_1))\}.$$

This is called Bellman's equation; the proof is very easy.

(2.5) Use the previous two exercises to show that the first-year optimality rule (2.11) holds also when $N = 3$. (You can proceed to the general case by induction, if you wish. A more direct proof will be given in the next section, however.)

3. COST EFFECTS

In the previous section we saw that time discounting could have the effect of making biological overexploitation appear profitable, possibly even to the point of extinction. Although many populations have been commercially exploited to a level close to extinction, the number of actual extinctions is still rather small.

As a population becomes reduced by harvesting, it is clear that the unit cost of harvesting is likely to increase. For example, consider the population of fish in a certain lake. The cost of catching one fish may be considered as approximately proportional to the time required to catch it.

Assuming that the rate of catch is proportional to the average density of fish in the lake, we see that the time required per catch is approximately inversely proportional to the number of fish in the lake. Letting $C(x)$ denote this unit harvesting cost, we then have

$$C(x) = \frac{c}{x}, \quad c = \text{constant} \tag{3.1}$$

As $x \to 0$, this unit harvesting cost becomes infinity. If p denotes the unit price of fish,* it is clear that fishing will not be "profitable" unless

$$p > C(x). \tag{3.2}$$

Let us define the zero-profit population level x_∞ by

$$p = C(x_\infty). \tag{3.3}$$

Then we conclude that the rational fisherman will never reduce the population below x_∞. Since $x_\infty > 0$, the fish population will not be harvested to extinction.

Several weaknesses are evident in this "non-extinction" argument. First, the actual number of fish must be an integer, so that whenever $x < 1$ the population is ipso facto extinct. Thus if $x_\infty < 1$, extinction is economically feasible. Indeed, since fish reproduce sexually, extinction is feasible if $x_\infty < 2$. More generally, there may be a "minimum viable population level" K_1 such that extinction results (or is highly probable) whenever x falls below K_1.

Secondly, the inverse proportionality given in Equation (3.1) may be unrealistic in many cases, since it is based on an assumption of uniform density of the fish population. Also the assumption that the time per unit catch is inversely proportional to the density involves a tacit assumption of random search. If the fish can be detected visually, this assumption is not valid. In such cases, the cost of extinction may be far from infinite.

In the sequel, we shall usually drop the specific form (3.1) for $C(x)$. and merely assume that

$$C(x) \text{ is a continuous nonincreasing function of } x \ (x > 0). \tag{3.4}$$

---

* In the case of sports fishing, p might represent the "value" of a fish to the fisherman.

Thus $C(0)$ may be either finite or infinite. Since $C(x)$ represents the cost of a unit harvest, which reduces the population from x to $x - 1$, the total cost of a harvest h that reduces the population from x to $x - h$ equals*

$$TC(h, x) = C(x) + C(x - 1) + \ldots + C(x - h + 1)$$

$$\approx \int_{x-h}^{x} C(y)\, dy. \tag{3.5}$$

We now incorporate this cost model into our harvesting model. From Section 2 we have

$$x_{k+1} = F(x_k - h_k); \quad x_1 \text{ given} \tag{3.6}$$

$$0 \leq h_k \leq x_k. \tag{3.7}$$

The net revenue from the harvest h_k is

$$\pi_k = ph_k - TC(h_k, x_k) = \int_{x_k-h_k}^{x_k} \{p - C(x)\}\, dx. \tag{3.8}$$

Hence the present value of all revenues equals

$$P.V. = \sum_{k=1}^{\infty} \alpha^{k-1}\pi_k \tag{3.9}$$

where $\alpha = \frac{1}{1 + i}$.

The problem of determining the optimal harvest policy $\{h_k^*\}$ will be solved by the well-known technique of "educated guessing."** Namely, we will guess that the optimal policy has the same characteristic as in Section 2, in the sense that there exists an optimal equilibrium escapement population $x = q$, and the optimal policy consists of reaching this level as rapidly as possible. All we have to do is to determine q. The correctness of this solution will be proved rigorously later.

---

* This involves a tacit "additivity" assumption regarding the cost of fishing.

** Guesses are of two kinds--lucky and unlucky. Educated guesses are more likely to be lucky than uneducated guesses, perhaps.

Assume that $x_1 > q$. Then we will have

$$h_1 = x_1 - q$$

$$h_2 = h_3 = \dots = F(q) - q.$$

Consequently

$$\begin{aligned} \text{R.V.} &= p(x_1 - q) - TC(x_1-q,\ x_1) + \sum_{k=2}^{\infty} a^{k-1}\{p[F(q)-q] - TC(F(q)-q,\ F(q))\} \\ &= p(x_1 - q) - TC(x_1-q,\ x_1) + \frac{\alpha}{1-\alpha}\{p[F(q)-q] - TC(F(q)-q,\ F(q))\}. \end{aligned}$$

Except for corner solutions, we require

$$\frac{d\ \text{P.V.}}{dq} = 0.$$

By elementary calculus, this becomes*

$$-p + C(q) + \frac{\alpha}{1-\alpha}\{p[F'(q) - 1] - [C(F(q))F'(q) - C(q)]\} = 0,$$

or, since $\alpha/(1-\alpha)$ reduces to $1/i$,

$$\{p - C(F(q))\}\ F'(q) - \{p - C(q)\} = i\ \{p - C(q)\}.$$

In order to simplify this equation, let us consider the sustainable economic yield** $Y(q)$ corresponding to the escapement level q, which is given by

$$Y(q) = p[F(q) - q] - TC(F(q)-q,\ F(q)) = \int_q^{F(q)} \{p - C(x)\}\ dx.$$

It is easy to see that the above equation can then be written as

$$Y'(q) = i[p - C(q)]. \qquad (3.10)$$

This (we hope!) is the solution to our problem. We may note first of all that (3.10) generalizes our previous result (2.13), for if $C(x) \equiv 0$ the two formulas are the same.

Equation (3.10) also has the same interpretation in terms of marginal productivity as (2.13). For consider the effect of a marginal increase $\Delta h = 1$

---

* Note, for example, that $\frac{d}{dq} TC(x_1-q,\ x_1) = \frac{d}{dq}\int_q^{x_1} C(y)\ dy = -C(q)$.

** Economists refer to $Y(q)$ as the economic rent.

in the first-year harvest, resulting in a corresponding marginal decrease in q. The immediate benefit will equal $[p - C(q)]\Delta h = p - C(q)$, which is equivalent to an annual revenue equal to $i[p - C(q)]$. The marginal loss in sustained economic revenue will equal $Y'(q)$. At the optimal escapement level q, these two effects must just balance, and that is what Equation (3.10) asserts.

What are the implications of Equation (3.10)? In particular, how does the optimal escapement level depend on the economic parameters i, p and $C(x)$? We must also ask whether the Equation actually has a solution, and if so, whether the solution is unique.

To answer these questions, it is convenient to rewrite (3.10) in the form

$$\frac{F'(q)}{1 + i} = \frac{p - C(q)}{p - C(F(q))}. \tag{3.11}$$

Let x_∞ be defined as in (3.3):

$$p = C(x_\infty).$$

Suppose first that such an $x_\infty > 0$ does exist. If $x_\infty \geqq K$ we know that harvesting is not profitable; hence we can assume that

$$x_\infty < K.$$

We can then see that (3.11) must have at least one solution $q = q_i$, because at $q = K$ we have LHS $< 1 =$ RHS, whereas at $q = x_\infty$ we have LHS $> 0 =$ RHS. By continuity, a solution q_i is assured (see Figure 4).

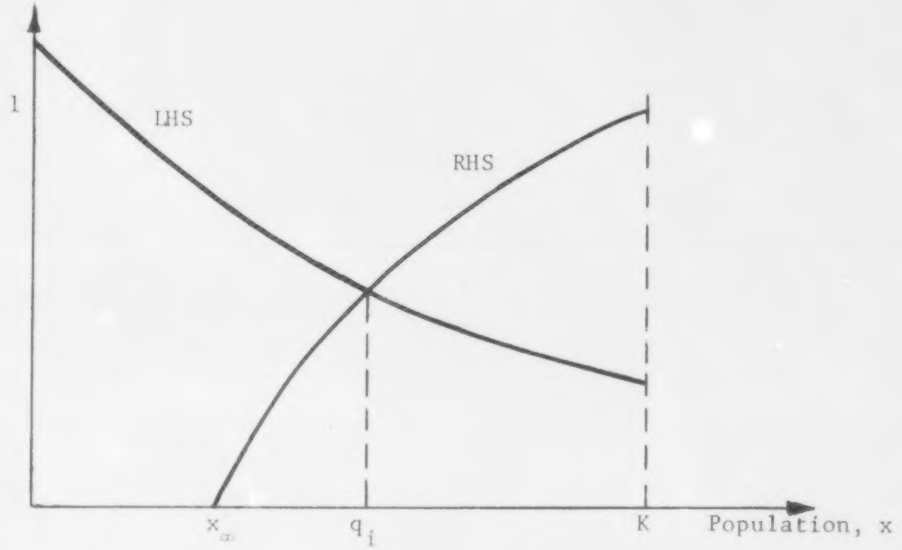


Figure 4

Unfortunately there appear to be no simple conditions which imply that the optimal escapement q_i is uniquely determined, although this is the case when $i = 0$ (see Exercise 3.1). On the other hand, numerical examples with more than one solution appear difficult to construct. Of course, our "guessing" technique is not very helpful when it gives more than one answer. And we shall see, when we come to the rigorous proof, that uniqueness of q_i is essential there.

Henceforth, therefore, we shall restrict attention to the case in which q_i is unique.

In this case we can deduce immediately that:

(i) q_i is a decreasing function of i;

(ii) $q_o > x_{MSY}$;

(iii) q_i approaches x_∞ as $i \to +\infty$.

Assertion (i) follows (Figure 4) from the facts that $F'(q)$ is decreasing, and that RHS < LHS for $q < q_i$. Assertion (ii) is true because $F'(q_o) < 1 = F'(x_{MSY})$. Assertion (iii) is obvious from Figure 4.

The implications of these results are as follows: (i) says that as the interest rate i rises* the optimal population level q_i falls; in other words, high interest rates lead to reduced "conservation" of resource stocks. This conclusion is perhaps the very essence of the idea of conservation: the less we value the future, the less we will be concerned with conservation. The conclusion also seems to be of practical importance, since in many cases (see Section 4) the level of conservation depends very sharply on the interest rate.

Whether the level q_i corresponds to biological overexploitation (i.e. whether $q_i < x_{MSY}$, or equivalently $F'(q_i) > 1$) depends upon the price and cost parameters, as reflected in the value of x_∞. If $x_\infty < x_{MSY}$, then (iii) shows that biological overexploitation will result from the use of sufficiently high interest rates. On the other hand, (ii) shows that overexploitation will never result from optimal harvesting at a zero rate of interest--and hence, also not at a sufficiently small rate.

---

* Once again, remember that this is the real opportunity--cost rate, not an inflationary rate.

Our model thus suggests that interest rates may be extremely important in conservation problems. Of course this is hardly surprising; everyone realizes that high interest rates (real or inflationary) can have serious repercussions in many aspects of the economy.

There is, however, another aspect of resource conservation that we have not mentioned. This is the problem of unregulated, common-property resources--an extremely important class of resources that includes most fisheries and wildlife populations, as well as the atmosphere and the oceans themselves.

Common-property resources

The reader is referred to the famous article "The Tragedy of the Commons," by Garrett Hardin (1968) for a detailed discussion of this problem. For the case of fisheries, the common-property problem was described earlier by Gordon (1954), whose model is related to ours. Very roughly, Gordon's analysis proceeds as follows.

Consider a seasonal fishery as described by our model. Let p denote the price of fish (assume constant) and let $C(x)$ denote the unit harvesting cost experienced by the individual fisherman. Whenever $C(x) < p$ the fisherman can make an immediate profit by catching more fish. Question: at what population level will fishing cease?

The answer, almost certainly, is that fishing will cease when $C(x) = p$, i.e. when $x = x_\infty$. First of all, when $x < x_\infty$ the fishermen are losing money with every additional fish caught--a situation that is not likely to continue for long. On the other hand, when $x > x_\infty$ any single fisherman who quits fishing simply allows the other fishermen to increase their profits. Unless all fishermen can agree to stop at some level $x_1 > x_\infty$ --and unless they have some means of enforcing this decision--they will be bound by their own self-interest to continue fishing.

The common-property fishery, therefore, behaves as if there were an infinite rate of interest, and hence it reaches a level of minimum conservation. In Hardin's words, without "mutual coercion mutually agreed upon," conservation is impossible.

Notice however that the common-property case does not necessarily result in biological overfishing--only if $x_\infty > x_{MSY}$ will this be true. If the price is low, or if costs are high, we will have $x_\infty > x_{MSY}$. Indeed, many fish populations are not exploited commercially at all, for the simple reason that

$C(x) > p$ for all population levels x. Most commercial fisheries have gone through stages of rising demand (increasing p) and increasing efficiency (decreasing $C(x)$). Some fisheries have been exploited for centuries without any overt sign of serious overfishing. It is only within the past decade or so that the overfishing problem has become really severe in some cases.

If the cost-price ratio is sufficiently low, i.e. if

$$C(0) < p,$$

we then have $x_\infty = 0$, and common-property exploitation will tend to the extinction of the biological resource. For the present-value maximization problem, the graphical solution of equation (3.11) now appears as shown in Figure 5. It is clear from this figure that $q_i = 0$ if i is sufficiently large. (See Exercise 3.2.) Thus even when unit costs of harvesting rise as the population level falls, present-value maximization could conceivably lead to extinction of the population. It might "pay" the owner of such a resource to destroy the resource and devote the proceeds to some alternative investment. In the next section we shall consider a case study (whaling) in which this conclusion appears to have some relevance.

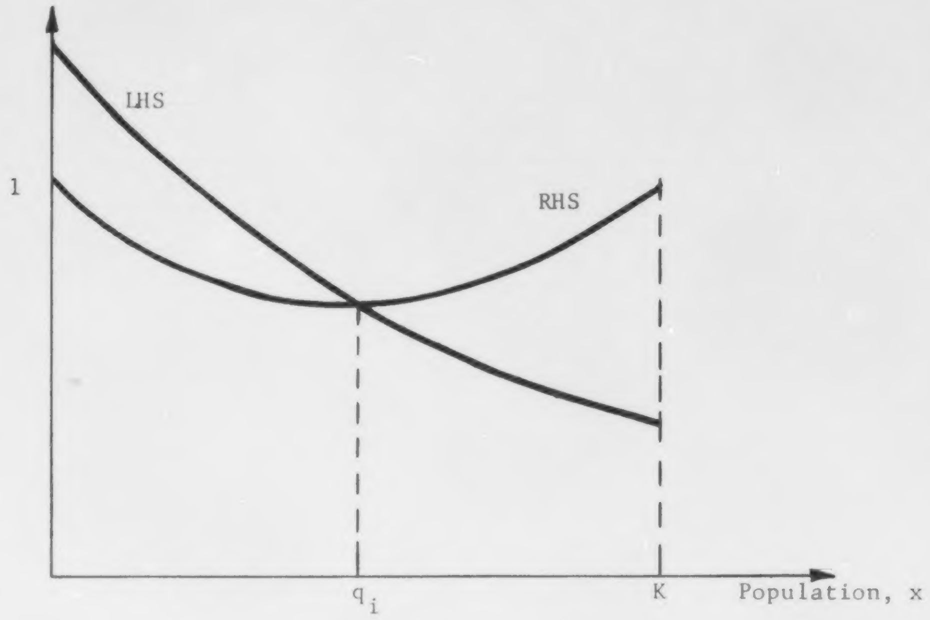


Figure 5

Proof of Optimality

The following simple but elegant proof that our "general" solution is correct, is due to M. Spence (1973).

We assume that Equation (3.11) has a unique solution $q = q_i > 0$. Then (see Figure 4) we have

$$\frac{F'(q)}{1 + i} \lessgtr \frac{p - C(q)}{p - C(F(q))} \quad \text{if} \quad q \gtrless q_i \quad \text{respectively.} \tag{3.12}$$

Since $F'(x) > 0$ the function $F(x)$ has a unique inverse function $\psi = F^{-1}$ defined also for $0 \leq x \leq K$. Let $\{h_k\}$ denote an arbitrary feasible harvest sequence and $\{x_k\}$ the corresponding sequence of recruitment levels, i.e.

$$x_{k+1} = F(x_k - h_k). \quad (x_1 \text{ given})$$

Then

$$x_k - h_k = \psi(x_{k+1}), \tag{3.13}$$

and by (3.8) and (3.5)

$$\pi_k = p\{x_k - \psi(x_{k+1})\} - \{G(x_k) - G(\psi(x_{k+1}))\} \tag{3.14}$$

where $G(x) = \int C(x)\,dx$ is an integral of $C(x)$.

Consider the present-value sum (3.9):

$$\begin{aligned} \text{P.V.}(\{h_k\}) &= \sum_{k=1}^{\infty} \frac{\pi_k}{(1 + i)^{k-1}} \\ &= \sum_{k=1}^{\infty} \frac{px_k - G(x_k)}{(1 + i)^{k-1}} - \sum_{k=2}^{\infty} \frac{p\psi(x_k) - G(\psi(x_k))}{(1 + i)^k} \\ &= \sum_{k=2}^{\infty} \frac{V(x_k)}{(1 + i)^{k-1}} + \{px_1 - G(x_1)\} \end{aligned} \tag{3.15}$$

where

$$V(x) = px - G(x) - (1 + i)\,\{p\psi(x) - G(\psi(x))\}. \tag{3.16}$$

Then by (3.12) we have

$$V'(x) = p - C(x) - (1 + i)\,\frac{p - C(\psi(x))}{F'(\psi(x))} \tag{3.17}$$

$$\gtrless 0 \quad \text{for} \quad x \lessgtr F(q_i) \quad \text{respectively.}$$

Now let $\{h_k^*\}$ denote the "most-rapid approach" harvest policy, which we guessed would be optimal. Thus

$$h_k^* = \begin{cases} x_k - q_i & \text{if } x_k > q_i \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

We are going to show by a direct calculation that the present value produced by $\{h_k^*\}$ exceeds the present value produced by any other feasible harvest policy $\{h_k\}$, i.e.

$$\text{P.V.}(\{h_k^*\}) > \text{P.V.}(\{h_k\}). \tag{3.19}$$

For $0 \leq s \leq 1$ define

$$w_k(s) = sx_k^* + (1 - s)x_k \tag{3.20}$$

where $\{x_k^*\}$ is the recruitment sequence resulting from the harvest policy $\{h_k^*\}$. Also define

$$I(s) = \sum_{k=2}^{\infty} \frac{V(w_k(s))}{(1 + i)^{k-1}}. \tag{3.21}$$

From (3.15) we have

$$\text{P.V.}(\{h_k^*\}) - \text{P.V.}(\{h_k\}) = I(1) - I(0). \tag{3.22}$$

To prove (3.19) we wish to show that $I(0) < I(1)$; this will be accomplished by showing that

$$I'(s) > 0 \quad \text{for} \quad 0 \leq s \leq 1. \tag{3.23}$$

In fact we have

$$I'(s) = \sum_{k=2}^{\infty} \frac{V'(w_k(s))(x_k^* - x_k)}{(1 + i)^{k-1}}. \tag{3.24}$$

Suppose first that $x_1 \geqq q_i$. Then $x_k^* = F(q_i)$ for all $k \geqq 2$. Whenever $x_k > x_k^*$ we have

$$x_k \geqq w_k(s) \geqq x_k^* = F(q_i)$$

so that by (3.17), $V'(w_k(s)) < 0$. Thus the corresponding summand in (3.24) is positive. Similarly $x_k < x_k^*$ implies positivity of the k-th summand. Hence $I'(s) > 0$ as required.

In case $x_1 < q$ we have $x_2^* = F(x_1) < F(q)$, and also (by feasibility), $x_2 \leqq x_2^*$. Hence

$$x_2 \leqq w_2(s) \leqq x_2^* < F(q)$$

so that $V'(w_2(s)) \geqq 0$. By continuing this line of argument, we conclude that $I'(s) > 0$ in this case also. QED

Exercises

(3.1) Besides (3.4), assume that $|C'(x)|$ is a decreasing function of x. Show that the ratio

$$\frac{p - C(x)}{p - C(F(x))}$$

is increasing when $x > \max(x_\infty, x_{MSY})$. Hence conclude that for $i = 0$ the optimal escapement q_o is unique.

(3.2) If either $p < C(0)$ or $1+i < F'(0)$, we know that $q_i > 0$, so that extinction is not optimal. Show that if $p \geqq C(0)$ and $1+i > [F'(0)]^2$ then extinction is optimal. Assume that $|C'(x)|$ is decreasing. [Hint. It suffices to consider $p = C(0)$. Use the generalized mean value theorem to show that, in this case,

$$\frac{p - C(F(x))}{p - C(x)} \cdot F'(x) < 1 + i \quad \text{for all } x.]$$

(3.3) Consider the following modifications (a)-(c) of the model discussed in this section. Determine the optimal harvest policy for one of the modifications, and give a rigorous proof.

(a) Finite time-horizon: Maximize $\sum_{k=1}^{N} \alpha^{k-1}\{ph_k - TC(h_k, x_k)\}$, subject to the usual conditions.

(b) Limited harvesting capacity: $\{h_k\}$ satisfies $h_k \leqq H$ = constant, as well as (3.7).

(c) Absolute conservation standard: $\{x_k\}$ is required to satisfy $x_k \geqq X$ = constant > 0.

4. APPLICATION: THE ANTARCTIC BLUE WHALE POPULATION

A well-known characteristic of cost-benefit analysis is the sensitivity of decisions to the rate of interest used. As a typical example, consider a dam project, in which the initial costs are large and where the full benefits will not accrue until many years after the completion of the dam. Only if the benefits are discounted at a sufficiently low rate will the net present value of the project be positive. Or to put it another way, only those projects that earn a sufficiently high rate of return will be acceptable.

The models described in the preceding pages suggest that conservation of biological resources may also be highly sensitive to the rate of interest. The marginal-productivity rule

$$Y'(q) = i[p - C(q)]$$

moreover, suggests that this sensitivity will be greatest for biological resources which have low marginal productivities Y'.

In forestry management, for example, the interest rate is notoriously crucial. Most temperate-zone forests grow only at 4-5 per cent per annum at best. When interest rates significantly higher* than this are used, forestry "management" reduces essentially to the clearcutting of all profitable stands.

In this section we shall apply our resource-harvesting model to another slow-growing resource, the Antarctic blue whale population. A brief review of the history of exploitation of this species will be given first.

The blue whale is the largest member of the family of baleen whales, all of which provide a commercially valuable, edible oil. In the early years of whaling, blue whales were simply too large to be captured. But with the development of motorized vessels, and especially of explosive harpoons, the blue whale became commercially valuable. In terms of our model, we may imagine that these technological developments reduced the cost function $C(x)$, so that eventually $C(K) < p$ was valid, and whaling became profitable.

In 1904 a whaling station was established on South Georgia Island, off the tip of South America, to process blue (and fin) whales. The main pelagic populations remained unexploited, however, since their feeding grounds, which covered some 10 million square miles of the Antarctic and South Pacific Oceans, were too remote for the land-based station.

---

* Remember that we refer to real interest rates, not inflationary rates; cf. Exercise 1.2.

In 1926, the Norwegian whalers developed a stern-slipway factory vessel capable of processing whales on the high seas--thus making $C(K) < p$ for the entire population. Following this development, whaling began in earnest (Figure 6), with 14 nations eventually owning and operating such factory vessels. In 1931 over 29,000 blue whales were harvested; this was more than 20% of the total Antarctic population--in one year!

By the 1950's it was clear that the blue whale was in trouble. The industry switched its attention primarily to the smaller, but still profitable, fin whale, continuing to take whatever blue whales were sighted.

The International Whaling Commission, which had been founded in 1946 to regulate the harvesting of all whale species, met annually but seemed incapable of stopping the overexploitation. The members of the Commission were, of course, the very nations involved in whaling.

It has sometimes been suggested that the Commission acted "foolishly" in failing to protect the whale stocks, thereby reducing the productivity of the industry. Although it may indeed be true that factors such as international jealousy and distrust may have contributed to the Commission's failure to protect the blue whales, it seems more likely that the commercial interests represented on the Commission simply recognized the adage "a bird [whale] in the hand [hold] is worth two in the bush [sea]." The whalers surely had positive discount rates--although the Russian whalers might be reluctant to admit it--so that biological overexploitation and optimal whaling may not have seemed contradictory to them.

In order to test this hypothesis, let us imagine the entire Antarctic whaling industry to be in the hands of a single profit-maximizing owner. We will consider the blue whale population only, in order to avoid complications of a two-species model. Using the model of Section 3, our first problem is to choose appropriate biological and economic components.

The biological growth function will be modeled by a quadratic expression*

$$F(x) = x + rx(1 - \frac{x}{K}). \qquad (4.1)$$

---

* This is related to the so-called logistic growth law used in theoretical biology.

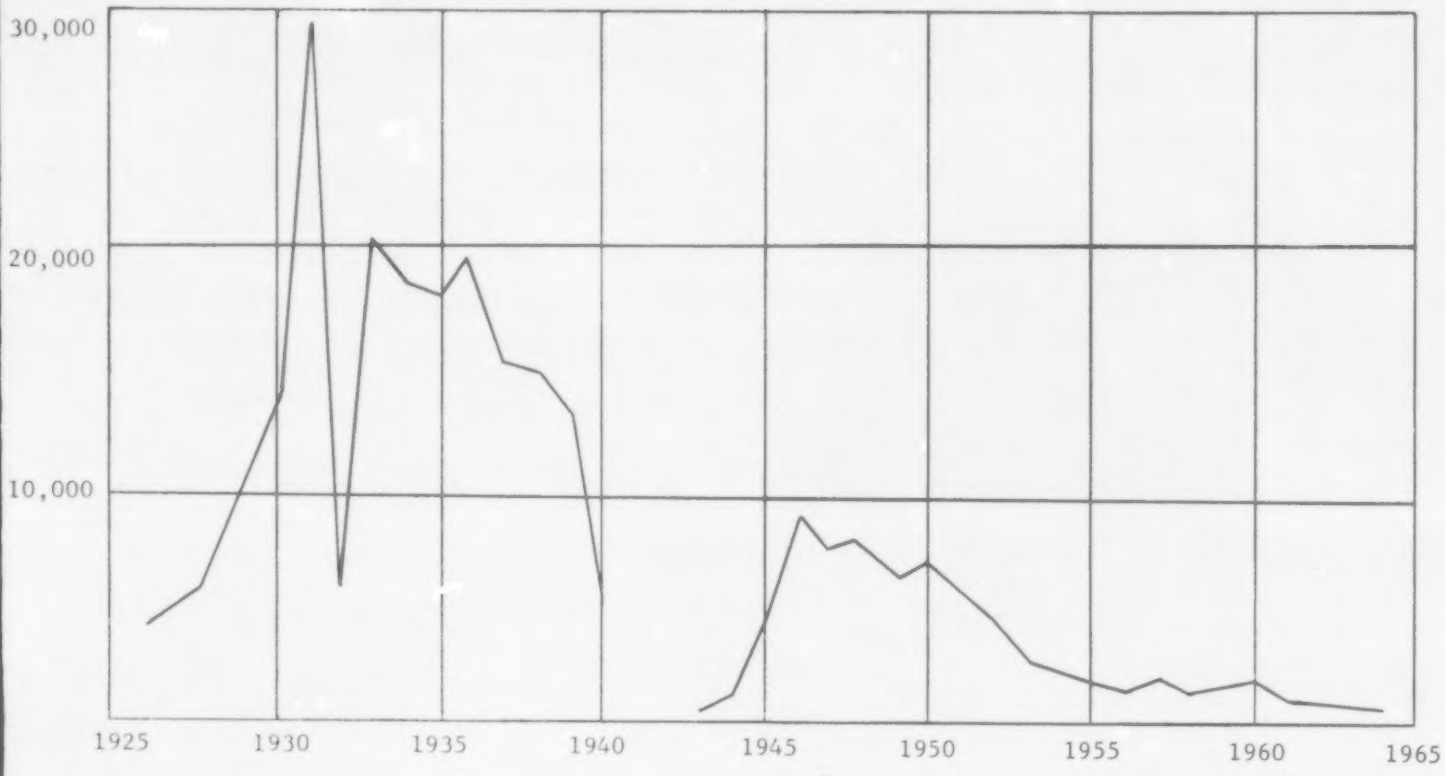


Figure 6. Antarctic Blue Whale harvests, 1926-1964

Recent estimates* of blue-whale population dynamics suggest that the maximum growth rate r may be about 5% per annum, and the equilibrium population K about 150,000 whales. We shall adopt these figures:

$$r = .05, \quad K = 150{,}000.$$

The market value of a blue whale is of the order of magnitude of \$10,000:

$$p = \$10{,}000$$

Next we shall suppose that the unit cost of whaling is inversely proportional to the population level:

$$C(x) = \frac{c}{x}. \tag{4.2}$$

The cost coefficient c is rather difficult to estimate, and also varies among the different whaling fleets. (Nations with high cost coefficients have already given up whaling.) The level x_∞ at which whaling becomes completely unprofitable, satisfies

$$x_\infty = \frac{c}{p}. \tag{4.3}$$

Thus c can be extimated from a knowledge of p and x_∞.

In 1965, when the whaling nations agreed to the protection of blue whales, the population level was around 5,000. Assuming (somewhat cynically) that such an agreement was only possible when the harvesting of blue whales was no longer significantly profitable, we may conclude that $x_\infty \geqq 5{,}000$. As mentioned earlier, however, the continued harvesting of the fin whale probably contributed to the decline of the blue whale population to a level below x_∞. Possibly 10,000, or even 20,000 would be a better estimate for the zero-profit level of blue whales alone. Let us agree on the conservative value

$$x_\infty = 20{,}000.$$

Except for the interest rate i, which we shall treat as a parameter, we now have everything needed to determine the optimal whale population, according to Equation (3.10):

---

* No "official" estimates of the blue-whale population dynamics, or of costs and prices, are available. The figures quoted here are rough estimates, mostly derived from personal correspondence with whaling scientists.

$$\frac{F'(q)[p - C(F(q))]}{p - C(q)} = 1 + i. \qquad (4.4)$$

The left-hand side of this equation was calculated numerically, and the resulting graph is shown in Figure 7.

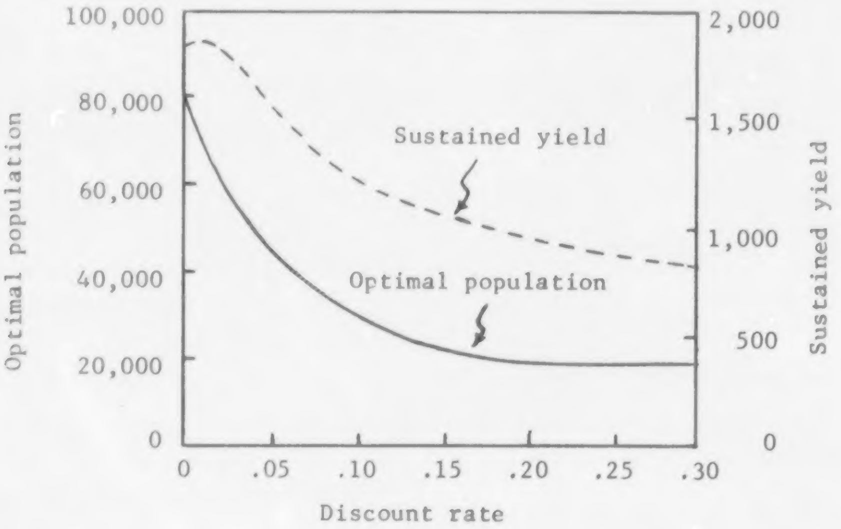


Figure 7

As expected, the influence of the interest rate is quite sharp. When $i = 0$ we have $q_0 = 85,000 > x_{MSY} = 75,000$. The corresponding sustained yield (shown by the dashed line) is $G(q_0) = 1842$ whales per year, compared to MSY = 1875 whales per year. For $i = 10\%$ the figures are $q_i = 31,000$ and $G(q_i) = 1235$ respectively; for $i = 20\%$ we have $q_i = 25,000$ and $G(q_i) = 1042$. We know from Section 2 that when $i > r = 5\%$, it is essentially only the costs of whaling that prohibit extinction from being "optimal."

This simplified model of whaling surely raises more questions than it answers. All sorts of complications, both biological and economic, have been overlooked. But the most difficult and important question, perhaps, centers around the very meaning of "optimality" for the management of resources such as whales. Suppose that the International Whaling Commission resolved to manage whaling in an optimal manner, for the benefit of humanity (not just for the whalers). What would it do? What interest rate would it adopt? Or would it use some criterion other than maximization of present value?

Economists tell us that present opportunity-cost rates of interest are far greater than 5% per annum.* Unfortunately, few people seem to recognize the effect that such high rates could have if they are used in resource-management decisions. The British Treasury Board, for example, has recently instructed government agencies to utilize a 10% discount rate. The end of forestry in Britain has been predicted as a result of this move--but there is no sign that Treasury Board officials anticipated such a result.

In the United States, the Sierra Club has been instrumental, along with economists, in getting Congress to require that a 10% rate be used on government construction programs. This has resulted in the non-approval of several proposed dam and irrigation proposals. But the same rate applied to management of the National Forests would lead almost surely to the complete clear-cutting of large areas. It seems doubtful that the Sierra Club would approve of such an outcome!

What then is the optimal discount rate to use in managing biological resources? A few pages of mathematics cannot be expected to provide the solution to such a difficult and far-reaching problem. Nevertheless mathematical models--*especially* simple strategical models of the kind described here--can contribute greatly to the understanding of such questions. The inter-relationships between economics and biology are far from obvious. The existing literature on renewable resource management is notable for its failure to disentangle these relationships properly. There are many unnecessary controversies, some of which have persisted for centuries (see Scott, 1972).

Any scientific field, pure or applied, needs a well-developed theoretical basis. Clearly dynamic optimization models are essential to any theory of renewable resource management.

Exercises

(4..) Recent reports on the Antarctic fin whale population indicate that the following piecewise-linear function may be a better approximation to the growth curve than the quadratic function used above:

$$F(x) = \max\ [(1 + r)x, K\}.$$

---

* The World Bank now recommends a 12% *real* interest rate for use in cost-benefit analysis by applicants for loans!

Estimated values of the parameters are $r = .08$, $K = 400,000$. If $x_\infty = 40,000$ is the zero-profit level for fin whales, determine q_i (numerically) in terms of the interest rate i.

(4.2) In this exercise we assume quadratic growth functions for both blue whales (x) and fin whales (y):

$$x_{k+1} = F(x_k) = x_k + rx_k(1 - \frac{x_k}{K})$$

$$y_{k+1} = G(y_k) = y_k + sy_k(1 - \frac{y_k}{L}),$$

where

$$r = .05, \quad s = .08, \quad K = 150,000, \quad L = 400,000.$$

Suppose that a constant annual "effort" E is devoted to the combined blue-fin whale industry, so that the harvest of each species is proportional to Ex and Ey respectively:

$$x_{k+1} = F(x_k) - aEx_k$$

$$y_{k+1} = G(y_k) - aEy_k$$

where a = constant. Show that x_k and y_k will approach equilibrium populations $\bar{x}$ and $\bar{y}$ satisfying

$$r(1 - \frac{\bar{x}}{K}) = s(1 - \frac{\bar{y}}{L})$$

provided that $\bar{x} > 0$ and $\bar{y} > 0$.

In this case we must have

$$\bar{y} \geqq L(1 - \frac{r}{s}) = 150,000.$$

What happens if, in fact $\bar{y} < 150,000$?

5. EVALUATION OF THE MODEL

Compared to the complexity of real-world problems in biology and economics, the model discussed in these notes is almost pathetically simple. Good strategic models, of course, should be kept as simple as possible. But on the other hand, it is essential to recognize explicitly the hypotheses underlying any mathematical model--simple or complex. In economics, for example, important policy recommendations are frequently based on specific models. If these models are inappropriate, the recommendations may be unwise. As in any scientific field, the development of economic theory proceeds by a continuous process of refining, improving, and evaluating existing models.

Let us list the major hypotheses underlying our model of optimal resource harvesting.

First, the model is a deterministic one, which also assumes perfect information, both on the biological and economic side. In real life, on the other hand, random variations and uncertainty are the rule. The optimal management of biological resources must obviously take these facts into account.

Second, the model assumes that the biological population can be represented by a simple parameter (x). In practice, the individual members of a population differ according to age, weight, sex, and also according to their spatial position at any given time. It is well known that a homogeneous model such as ours can result in misleading descriptions of population dynamics.

Third, the model utilizes many severely restrictive economic hypotheses. Price is assumed constant, unit harvesting cost is assumed to depend only on the population levels and not on the rate of harvest, etc. If such assumptions are highly unrealistic, the model can again be expected to yield unreliable results.

Finally, we have made numerous technical assumptions regarding smoothness and convexity of certain functions and the unique solvability of equations. These assumptions, too, may have serious repercussions.

Clearly our model is drastically oversimplified. Yet it may serve a useful purpose. Not, perhaps, the practical purpose of showing "how to" manage a certain resource, but the useful purpose nevertheless, of explaining some of the complex ways in which economics and biology may interact. Even computer studies that show, as they often do, the strong influence of interest rates on sustainable yield--even such studies do not help much in understanding this phenomenon. Such understanding (of a widely misunderstood phenomenon) can only come, it seems to me, from a simple model. If my years of teaching mathematics have taught me anything, it is that techniques are easy to come by, but understanding is very difficult.

NOTES AND REFERENCES

Only a few references to the literature will be given here. Further references may be located in the bibliographies of the listed references, as well as in my forthcoming book mentioned in the introduction.

Section 1. An elementary introduction to cost-benefit analysis is Mishan (1971). Hirshleifer (1970) is recommended for its treatment of the theory of interest and investment.

Section 2. Population models of the kind used here are discussed by Watt (1968) and Maynard Smith (1968). The definitive work on dynamic programming is Bellman (1957). Concerning the rate of discount, see Feldstein (1964).

Section 3. This section follows Clark (1973a); a related continuous-time model and various extensions are given in Clark (1973b) and Clark and Munro (1974).

Section 4. An interesting description of the blue whale industry is given by Small (1971). Statistics and scientific reports on whales and whaling are published regularly in the Annual Reports of the International Whaling Commission, available at libraries which are U. N. depositories. The economics of the joint blue-fin whale industry is discussed by Clark (1974).

Bellman, Richard. Dynamic Programming. Princeton University Press, 1957.

Clark, Colin W. "Profit maximization and the extinction of animal species." Journal of Political Economy 81 (1973), 950-961. 1973a.

Clark, Colin W. "The economics of overexploitation." Science 181 (14 Aug), 630-634. 1973b.

Clark, Colin W. "Antarctic whaling: a model of joint production." Manuscript, 1974.

Clark, Colin W. and Munro, Gordon R. "The economics of fisheries and modern capital theory: a simplified approach." Manuscript, 1974.

Feldstein, M. S. "The social time-preference discount rate in cost-benefit analysis." Economic Journal 74 (1974), 360-379.

Hardin, G. "The tragedy of the commons." Science 1962, 1243-1247, 1968.

Hirshleifer, J. *Investment, Interest and Capital*. Prentice-Hall, 1970.

Maynard Smith, J. *Mathematical Models in Biology*. Cambridge University Press, 1968.

Mishan, E. J. *Cost-Benefit Analysis*. Union, 1971.

Scott, Anthony D. *Natural Resources: The Economics of Conservation*, 2nd ed. McLelland-Stewart, 1972.

Small, George. *The Blue Whale*. Columbia University Press, 1971.

Spence, Michael. "Blue whales and optimal control theory." Institute for Mathematics Studies in the Social Sciences, Report No. 108, Stanford University, 1973.

Watt, K. E. F. *Ecology and Resource Management*. McGraw-Hill Book Company, 1969.

Chapter 6

SOME EXAMPLES OF MATHEMATICAL MODELS FOR THE DYNAMICS OF SEVERAL-SPECIES ECOSYSTEMS

H. R. van der Vaart
Institute of Statistics
North Carolina State University

0. PREFACE

The mathematics that is applied in this module is at several levels. First, calculus is assumed throughout; especially the ability of finding primitive functions will come in handy. Second, extensive use is made of the isocline-and-arrow method; the background for this method is given in appendix A, sections A.1 and A.2. Third, for the study of differential equations in the neighborhood of singular points the linearization method is used; this method is presented at length in appendix A, sections A.3 and A.4. Finally, section 1.6 investigates certain aspects of difference equations; most of the needed mathematics is developed in that section; see the quoted literature for more information.

One can skip certain sections with relative ease, e.g., section 1.6 or section 2.3, but the instructor will be able to spot other possibilities as the work progresses. One can also skip some of the questions that require more intricate mathematical work. One can choose between going over sections A.3 and A.4 in detail, or just picking out the results. Finally, it should be said that our presentation proceeds in an analytic and geometric framework, but that the experience of past users indicates that it is quite effective to supplement this approach by work on analog or digital computers, provided adequate guidance is available.

Depending on which of the above choices one makes, one may spend anywhere between four and ten weeks on this module.

Besides the appendix A on certain mathematical tools, please note appendix B (the bibliography) and appendix C (the teacher's manual).

1. INTRODUCTION

1.1. Background.

The natural interaction of organisms and their environment is called an "ecosystem," a term introduced by Tansley (1935). "Biogeocoenosis," a term coined by Sukachev at about the same time, is perhaps more suggestive. It depicts a more or less complex interaction between biological community (biocoenosis) and habitat (biotope, expressed in the above word by the syllable "geo").

"The individual ecosystems vary greatly in size and structure. The entire globe is an ecosystem, the only one which is not influenced by other ecosystems. An island, a forest, a pasture, a decaying tree stump with its moss and fungi, even a puddle on the path which is only temporarily inhabited, all such natural phenomena deserve to be called ecosystems. Thus, great variations exist not only in magnitude, duration and production, but also in the degree of dependence on other ecosystems." (quoted from p. 2 of H. Ellenberg, 1971).

It should be emphasized (cf. p. 26 of J. Balogh, 1958) that neither biocoenoses nor biotopes exist on their own in nature, and that they are separated in our thinking only.

A long-standing effort in mathematical modeling of ecological problems concerns the variations in numbers of individuals belonging to the various species living together in an ecosystem or biogeocoenosis. The book by U. d'Ancona (1954) provides a good entry to the older literature in this field. We will discuss some examples of such mathematical models, assuming that no migration takes place into or out of the ecosystem (that is, the theory is restricted to the case of no migration). We will also assume that the ecosystem has fixed spatial boundaries (as more or less suggested by Ellenberg's above-quoted examples), so that we may equivalently describe events in terms of total numbers of individuals in the system or in terms of density (number of individuals per area or volume, as the case may be). There is obvious practical interest in the number aspects of population dynamics; think of insect pests, fishery research, population explosion.

Between any two species in an ecosystem any one of several relations may exist, such as:

a) prey-predator (first species is eaten by second species), and host-parasite ("a small organism lives in or with and at the expense of a larger animal or plant", cf. p. 253 or W. C. Allee et al., 1949);

b) cooperation ("helpful interactions between organisms," W. C. Allee et al., 1949, p. 395; or "mutually beneficial relationships between species," p. 710 of same), called mutualism by some authors, symbiosis by others;

c) disoperation (harmful interactions between the organisms of the same or different species, cf. W. C. Allee et al., 1949, p. 395), e.g., most cases of competition where both want the same things from their surroundings.

1.2. The historical beginnings of mathematical models in ecosystem dynamics.

If one should be asked to provide the name of one scientist most intimately connected with the origin and development of the mathematical models to be discussed, one probably ought to mention Vito Volterra. According to the very informative biography of V. Volterra, written by Sir Edmund Whittaker and published in 1941 in the "Obituary Notices of Fellows of the Royal Society," reprinted in the Dover paperback, Volterra (1959), pp. [5] through [39] (especially see pp. [19] through [25]), Volterra was interested in biological (and social) applications of mathematics as early as 1901. His active role in research in this area does not seem to have gotten underway before 1925, though, and was prompted by conversations with U. d'Ancona (also see d'Ancona, 1954, p. 52) on the fluctuations of fish caught in the Adriatic Sea. In d'Ancona's words, however, "from the beginning Volterra attacked the problem in its general aspect." Indeed, the only trace of this motivation for Volterra's work seems to be the fact that one of his earliest papers on the subject, viz. a memoir (Variazioni e fluttuazioni in specie animali conviventi) of 142 pages, was published in the Rendiconti Comitatio Talassografico Italiano (Memoria 131, Venezia, 1927), which according to its name was beholden to matters of the sea. It would be interesting to compare the results of Volterra's (or indeed, anybody's) theory with the data that prompted its development. In the bibliography of d'Ancona (1954), some publications (all in Italian) by d'Ancona are listed that would probably provide such information (see also Zupanović, 1953, quoted on p. 259 of d'Ancona, l.c.), but the author of the present exposition has not been in a position to consult these publications.

Two other authors should be mentioned here for their substantial contributions to the early development of this field: Alfred J. Lotka and G. F. Gause (not to be confused with the mathematician C. F. Gauss). Some of Lotka's work actually preceded Volterra's; see Lotka (1925, 1956), especially pp. 88-99 in Chapter 8, where some of his earlier work is also quoted. As for his motivation, it seems to have been more the desire for a general understanding of the "laws governing the distribution of matter among complexes of any specified kind" (see p. 216 of Lotka, 1907) than such practical matters as were brought before Volterra by d'Ancona. Chemical systems were very much on Lotka's mind: see P. 204 ('monomolecular reactions') and p. 216 ('We have recognized the problem of chemical dynamics as a special case of a wider problem') of Lotka (1907), which paper in its first half discusses the relations between birth and death rates, age distributions and life tables, problems of long-standing interest with students of human populations (demographers). Gause was a population ecologists who conducted quite a few laboratory experiments. From the titles of some of his papers of 1934 and 1935 (listed in the bibliography of his book, Gause, 1934 and 1971), as well as from the overall thrust of the same book, it becomes apparent that he was quite interested in experimentally exploring Volterra's theory (also see Gause, 1935). Among other things, he contributed greatly to the establishment of the idea that whenever two species compete against each other, one species often gains the upper hand completely (cf. sec. 3.3 of the present exposition).

1.3 A partial list of more recent literature on the subject.

From the first beginnings by Lotka and Volterra until the present time hundreds of papers have been published, either to further develop the mathematical theory, or to expound experimental work elicited by some phase of the mathematical theory (such as Gause's above-mentioned work), or to set forth non-mathematized theoretical reflections on matters arising in this area.

If anybody wants to enter this vast amount of literature, he can rely on several books and surveys. Besides the above-quoted books by Gause and d'Ancona (both available in paperback) there are several other possibilities. Some suggestions follow:*

J. Maynard Smith (1968), Mathematical ideas in biology; Cambridge University Press, Chapters 2 and 3; available in paperback, although only few references are quoted, these chapters give an exposition that might help some students.

J. Maynard Smith (1974), Models in ecology; Cambridge University Press; 146 pp.; emphasis on analysis of theoretical models; 87 references, some of which are concerned with observations.

M. Williamson (1972), The analysis of biological populations; London, Arnold & New York, Crane and Russak, 180 pp.; more emphasis on genetical aspects and somewhat less on mathematical treatment than the preceding books; 286 references.

---

* The first five titles on this list should not give too many mathematical difficulties to undergraduate math majors; the next two are more advanced; the last title is definitely at the graduate level. The fourth and the seventh title are biologically more demanding than the others.

D. B. Mertz (1972), The Tribolium model and the mathematics of population growth; pp. 51-78 in: Annual Review of Ecology and Systematics, vol. 3; 102 references. Emphasis mainly on non-deterministic observations and models. Tribolium is the flour beetle.

R. H. MacArthur (1972), Geographical ecology; patterns on the distribution of species; New York, Harper and Row, 269 pp.; Chapter 2 (37 pp.) on competition and predation.

R. M. May (1973a), Stability and complexity in model ecosystems; Princeton University Press, 235 pp.; available in paperback; circa 230 references, a majority of which is of a theoretical nature, with a strong minority on observational work.

M. B. Usher and M. H. Williamson (eds.)(1974), Ecological stability; London, Chapman and Hall, and New York, Wiley; 196 pp.; a collection of 12 papers on various related subjects, theoretical and practical; each paper carrying its own list of references (between one and two pages long).

N. S. Goel; S. C. Maitra and E. W. Montroll (1971), On the Volterra and other nonlinear models of interacting populations; Academic Press; A Review of Modern Physics Monograph; 145 pp.; 91 references; a group of physicists building a superstructure on the Volterra models (Reprint of: Reviews of Modern Physics 43, pp. 231-276, April 1971).

A high denisty of relevant papers can be found in some issues of such journals as The American Naturalists, Theoretical Population Biology, Acta Biotheoretica, Bulletin of Mathematical Biology (formerly: Bull. of Math. Biophysics), Journal of Theoretical Biology, Mathematical Biosciences, Ecology.

1.4. General remarks about the degree of realism and the mathematical level of the models to be discussed.

In view of the overwhelming volume of literature in our subject area, coupled with the fact that the audience of the present exposition consists of undergraduate students in mathematics, it is self-evident that we cannot hope to do much more than provide an outline of the early stages of this developing field. There is no way in these few pages to bring the reader up to the full spectrum of present-day research. Some of the newer developments use mathematical tools that are fairly close to the ones used in the present exposition, but most of these developments, having gone in the direction of greater realism, require either more sophisticated mathematics or big computer programs (hence some knowledge of more sophisticated numerical methods). Thus, because of lack of time as well as because of lack of background on the part of the audience, this exposition will have to stop far short of the satisfaction of modeling the world, or a forest, or the tundra, or in fact that "simple" puddle in your backyard.

There is, however, nothing shameful in retracing some of the steps that research mathematicians and biologists have gone through during the last few decades. Many undergraduate physics courses do not even come close to doing that sort of thing in physics, and ecosystems are rather complicated objects of research. The reader will reach a level at which he can recognize what many papers, e.g., in the American Naturalist, are concerned with, even if he will not (yet) be able to write those papers himself. (The reader will also learn a few mathematical techniques that are very helpful in studying the behavior of solutions of (especially 2-dimensional) systems of nonlinear ordinary differential equations of order one, which do occur in several areas of applied mathematics).

An aspect of mathematical model building, of which the reader should be advised, an aspect that is particularly obvious in the discussion of such many-faceted objects as ecosystems, is the following. The very first model that Volterra put up (for a predator-prey system), although it yielded some interesting results, was biologically trivial. The models that came after it cannot usually be solved analytically; they have to be studied by qualitiative methods or numerically, by means of a computer. Some people get discouraged when they realize that, in spite of all that work, such models are still at kindergarten level, biologically speaking. There is, of course, no reason for such discouragement. Research ecologists have actually worked with these models; these models have elicited a great deal of experimental work, either to refute or to expand them. This is the hallmark of the development of science: a continuing dialogue between experiments or observations on the one hand, and theoretical considerations on the other, with the final decision as to the "truth" being a long time in coming or even, perhaps more commonly, forever elusive. Experience shows that many mathematicians find this type of situation distasteful. Yet, whoever aspires to being a mathematician with applied interests in life or social sciences had better adjust to the idea that any mathematical structure that he may propose as a reflexion (as a model) of a real life situation may be criticized or refused by his (biological or social) partner, not because anything is wrong with his mathematics, but because his choice of model does not satisfy his partner on biological (or social) grounds.

In order that the reader may grow accustomed to this continual "threat" of criticism on the quality of whatever model he might be working with at a given moment, we have scattered discussion questions throughout the exposition. In these the reader is invited to explore the weak points of a model which has just been introduced or discussed in the main body of the exposition, to explain why the reader thinks certain aspects of the model are weak, and to set forth the way(s) in which the reader would seek improvement, thus generating alternative models. If adequate computer facilities are available, these discussion questions could be the source of projects in which a group of students would numerically explore some of these suggested "improvements." They could discover that indeed the solutions to the "improved" systems show more realistic features--or that they look worse than the original systems. These discussion questions are, in a way, open ended: they have no unique, correct answer; they serve to sharpen the critical sense of the reader, who will have to live by his wits, sailing an uncharted course.

In the face of this critical attitude that the present exposition seeks to develop in the reader, it is important on the other hand not to lose sight of the fact that all major models discussed in this exposition have actually been published in the research literature: none of them has been contrived for the special purpose of this project.

A final remark concerns the adduction of records of experiments or observations, so that the reader may have numbers available against which to test this or that model. Unfortunately, biological publishing habits do have a marked preference for graphs over tables. At the time these lines were being written, the author still had not located one publication in which actual numbers were given: the tables given on pages 58 through 60 of Leigh (1968) were read from certain graphs (as stated on pp. 53 and 56 of that article); it is better that the reader should glean his own numbers from these (or other) graphs. As for the problem of fitting any of the models under discussion to such data, the best the reader can do is eyeball them: optimal statistical methods for this purpose are the subject of active research, but a rough estimate can be obtained from looking at phase plane plots of the data.

1.5. Specific remarks about the mathematical nature of the models to be discussed.

Our discussion will start with examples of mathematical models for systems consisting of only two species (to be expanded to more than two species in some of the later sections). Such systems are rare in nature, but have been studied in the laboratory (e.g., Gause, 1934, and Huffaker, 1958). They have the pedagogical advantage that at least some of their mathematical properties can be handled by elementary means, because a two-dimensional space can be depicted easily on a piece of paper or on a blackboard. Thus they are accessible even to persons with a modest mathematical background. At the same time, in most real world situations one must look at more than two species (e.g., see the discussion around equation (4,5) below). Some of the mathematical tools needed for the study of these larger, more general systems can be introduced during the discussion of the two-species models, so that the audience may get acquainted with them. Other techniques will have to wait till later (i.e., till after this course). In any case nobody should get away with the impression that ecosystems of more than two species are merely the frosting on the cake. In most real ecosystems they are very much the cake itself. It is the introductory flavor of the present exposition that prompts a relative emphasis on two-species systems.

Another feature of our models, with respect to which they may be said to exhibit a certain lack of realism, is the fact that they are deterministic: these models do not provide a way of accounting for the many hard-to-control influences that are characteristic of most biological experimentation and whose effects are clear from, among other things, the graphs in the book by Gause (1934, 1971); for instance see his Figures 24 and 25 on p. 106 and 107. There are several reasons for our restriction to deterministic models. First, the originators of the theory did it, probably on the strength of the example set by the physical sciences; and countless papers have been, and still are being published in this vein. As was said before, there is no reason for an elementary exposition to be ashamed of introducing its audience to such vast resources of literature. Second, most of the stochastic models (note that the standard method of allowing a mathematical model to cope with observations that are not 'on the curve' is the incorporation of probabilistic features) are mathematically intractable, and so most of the work done on them has consisted of computer simulation (Monte Carlo) studies (e.g., Bartlett, 1961; Leslie and

Gower, 1960; Leslie, 1962); these studies took off from discrete models, see Question 1.6,6. Now, although this entire module does offer many opportunities for computer-oriented projects, we did not want to slant the main exposition in that direction. A look at some of the literature (e.g., Reuter, 1961; Iglehart, 1964; Tsokos and Hinkley, 1973) on mathematical and analytical work that has been done on stochastic models amply demonstrates that this work is too advanced for our present purposes. Third, the deterministic models are often heuristically very useful for suggesting problems to be studied by stochastic methods (cf. Feller, 1939; Neyman, Park and Scott, 1956; Mertz, 1972). All in all, these deterministic models are a good port of entry into this area, but again, the audience should not stay in this port for the rest of their professional lives.

Finally, the great majority of the following models have the form of differential equations. This point has three main implications: (α) the models predict the immediate future on the basis of the 'present', not on the basis of any part of the past, (β) the increment of any of the state variables over a short time interval of length h is assumed to be a rather special type of function of h, and (γ) the models are basically 'continuous,' rather than discrete. Let us discuss these points one at a time.

α) The models may be regarded as a limiting case of a subset of the following category of models. If $x(t)$ represents the number of individuals of species 1 at time t and if $y(t)$ represents the number of individuals of species 2 at time t, then the increase of both during the time elapsing between t and $t+h$ is assumed to depend on h, on $x(t)$, and on $y(t)$:

$$\left\{\begin{array}{l} x(t+h) - x(t) = F(h, x(t), y(t)) \\ y(t+h) - y(t) = G(h, x(t), y(t)) \end{array}\right\}. \tag{1.5,1}$$

Discussion question 1.5,1. Can you think of (a) reasons(s) why, or of (an) example(s) in which, it might not be a good idea to express these changes (viz., $x(t+h) - x(t)$, and $y(t+h) - y(t)$) during the time interval from t to $t+h$ solely in terms of $x(t)$ and $y(t)$? In discussing this question, possibly for a one-species population such as the human population, try to think of relations between increments like $x(t+h) - x(t)$ and the present as well as past values of the population size. Any example where the link to the past seems quantitatively important would provide a situation in which the type of model exhibited in equation (1.5,1) should be viewed with suspicion and replaced by a model in which $F(\ )$ and $G(\ )$ depend also on values of $x(\ )$ and $y(\ )$ at times $< t$.

Remark 1.5,1. On the other hand, models which do incorporate the past, such as

$$\begin{Bmatrix} x(t+h) - x(t) = F(h, x(t), y(t), x(t-\tau), y(t-\tau)) \\ y(t+h) - y(t) = G(h, x(t), y(t), x(t-\tau), y(t-\tau)) \end{Bmatrix},$$

are mathematically much more difficult than equation (1.5,1) (cf. question 1.5,10). Also, even in the case of human populations, where omission of the past is quite suspicious, certain models of the form of equation (1.5,1) have satisfied demographers such as Pearl and Reed for a long time. So on the basis of the same philosophy as expressed before, most of the discussions will refer to models of the form of equation (1.5,1).

β) It will then be assumed that the functions $F(h, x, y)$ and $F(h, x, y)$ of equation (1.5,1) are of the form $h\cdot\varphi(x, y)$ and $h\cdot\psi(x, y)$. So the equation (1.5,1) can be written in the form

$$\begin{Bmatrix} \dfrac{x(t+h) - x(t)}{h} = \varphi(x(t), y(t)) \\ \dfrac{y(t+h) - y(t)}{h} = \psi(x(t), y(t)) \end{Bmatrix}. \qquad (1.5,2)$$

Remark 1.5,2. This assumption is similar to a standard assumption made in all classical models of physics and chemistry. Yet, although it is natural to assume that $|x(t+h) - x(t)|$ is an increasing function of h, it is not clear why $|x(t+h) - x(t)|$ could not be proportional to $h^{1-\varepsilon}$ or to $h^{1+\varepsilon}$ $(\varepsilon > 0)$, to name just two examples. In other words, although physics and chemistry have conditioned us to think, talk and act in terms of rates of change per time unit (in fact, in terms of quantitites such as $[x(t+h) - x(t)]/h)$, yet the mere fact that we have been so conditioned does by no means prove that this type of quantity necessarily provides the right kind of framework for ecological theory.

Discussion question 1.5,2. Show that the next step in the modeling process (see (γ) below) would result in the same equation (1.5,3) if we would make the following weaker assumption about $F(h, x, y)$ and $G(h, x, y)$: $F(h, x, y) = h\cdot\varphi^*(h, x, y)$ and $G(h, x, y) = h\cdot\psi^*(h, x, y)$, provided $\lim\limits_{h\to 0} \varphi^{*}(h, x, y) = ?$ and $\lim\limits_{h\to 0} \psi^{*}(h, x, y) = ?$

$\underline{\gamma}$) Finally, again in analogy to physical and chemical models (a fine specimen of this kind of argument can be found in sections 8-2, 8-3 and 9-5 of volume 1 of the Feymann Lectures on Physics, 1965, where the traditional definition of speed is defended) it is argued that equation (1.5,2) should hold for any h, small or not-so-small, so that we may as well go all the way, let $h \to 0$ and cast our model in the form of differential equations,

$$\left\{\begin{aligned} \dot{x}(t) &= \varphi(x(t), y(t)) \\ \dot{y}(t) &= \psi(x(t), y(t)) \end{aligned}\right\}. \qquad (1.5,3)$$

Remark 1.5,3. This step is odious to many a biologist because his observations on populations are never made as in a time continuum, but at separate moments of time, and because 'number of individuals' (the entity represented by $x(t)$ and $y(t)$, see the above point (α)) is discrete and integer-valued. Mathematically speaking, the juxtaposition of discrete-time and continuous-time models is not as harmless as one might think (see section 1.6). Biologically speaking, one could argue that x and y may be thought of as representing the biomass* of these species, a 'continuous' quantity (although the observational technique does consist in counting (or estimating) the number of individuals); or as representing the mean value (again, a 'continuous' quantity) of discrete variables subject to (relatively small) random fluctuations superimposed on the broader sweep that is then described by the model of equation (1.5,3). In section 1.6 we will give a few examples of the problems surrounding the comparison of discrete-time and continuous-time models.

Once we have accepted the idea of differential equations as a category of models for ecosystems, the main question left with respect to model building is how to specify the functions $\varphi(\;)$ and $\psi(\;)$ in equation (1.5,3). We will begin with the observation that if species 1 would happen to be extinct at some moment in time, then $\dot{x}(t)$ must be zero from that moment on (no immigration!). Same for $\dot{y}(t)$. So

---

* Although students of population dynamics readily propose the concept of biomass as a way out of the dilemma between discrete reality and continuous model, I do not know of an actual implementation of this concept. One could for instance use live weight, or dry weight, as measures of biomass.

$$\left\{\begin{matrix} \varphi(0,y) = 0 \\ \psi(x,0) = 0 \end{matrix}\right\}. \tag{1.5,4}$$

Although this can be achieved in various ways, one traditionally satisfies equations (1.5,4) by putting $\varphi(x,y) = x\cdot f(x,y)$ and $\psi(x,y) = y\cdot g(x,y)$ (with $f(\ )$ and $g(\ )$ bounded). Thus equation (1.5,3) reduces to

$$\left\{\begin{matrix} \dot{x}(t) = x(t)\cdot f(x(t),\ y(t)) \\ \dot{y}(t) = y(t)\cdot g(x(t),\ y(t)) \end{matrix}\right\}, \tag{1.5,5}$$

or in abbreviated notation (remember that x and y are functions of t):

$$\left\{\begin{matrix} \dot{x} = x\cdot f(x,y) \\ \dot{y} = y\cdot g(x,y) \end{matrix}\right\}. \tag{1.4,5a}$$

In the following chapters several choices of $f(\ ,\ )$ and $g(\ ,\ )$ will be discussed, in accordance with the type of relations (prey-predator, cooperation, disoperation) obtaining between species 1 and 2.

Remark 1.5,4. The equation (1.5,4) is commonly, but not universally, adopted. Some publications discuss alternatives. In the sequel there will be discussion questions and/or remarks to draw attention to these exceptions.

As a first and very simple example of the principle expressed by equation (1.5,5), and also as preparation for further developments, let us take a very brief look at a model (in the spirit of equation (1.5,5)) for the growth of a population consisting of one species only (e.g., human populations), namely

$$\dot{x}(t) = x(t)\cdot f(x(t)), \quad \text{or} \quad \dot{x} = x\cdot f(x), \tag{1.5,5b}$$

where $f(x) = \frac{\dot{x}}{x}$ is the relative growth rate of the population. At an early date (in fact, after Malthus) demographers concluded that no (human) population was going to grow forever at a constant relative rate.

Question 1.5,3. Find $x(t)$ if $\dot{x}(t) = r\cdot x(t)$ for $t \geqq t_0$, $x(t) = x_0$, and r is a constant ('exponential' or 'logarithmic' growth).

So demographers argued that there should be some maximum population size, K say, that "the environment could carry" (K = "carrying capacity" of the environment) and that the closer the population size would come to K, the smaller $\dot{x}/x$ should be. So $f(x)$ in equation (1.5,5b) should be equal to 0 if $x = K$ and equal to a decreasing function of x for $0 < x < K$. One of the simplest such functions is $f(x) = \gamma\ (K-x)$, which reduces equation (1.5,5b) to

$$\dot{x}(t) = \gamma\ x(t)\ (K-x(t)), \quad \text{or} \quad \dot{x} = \gamma\ x\ (K-x), \tag{1.5,5c}$$

the so-called logistic equation (e.g., see Gause's little book, pp. 32-44).

Question 1.5,4. Show that the function

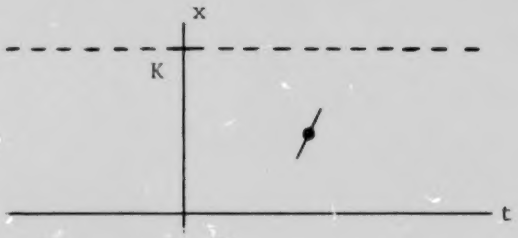
$$x(t) = \frac{K}{1 + \frac{K-x_o}{x_o} e^{-\gamma K(t-t_o)}} \tag{1.5,5d}$$

is a solution of equation (1.5,5c) and takes the value x_o at $t = t_o$ (hence is the solution of the logistic equation that satisfies the initial condition $x(t_o) = x_o$). This solution is also defined for $t < t_o$.

Discussion Question 1.5,5. From the form of the equation (1.5,5d) it would seem that this function of t depends of four parameters, γ, K, t_o, and x_o. Show that in reality there are only three parameters.

Question 1.5,6. For $\gamma = 1$, $K = 10$, carefully draw a graph of the function of t that is defined by equation (1.5,5d), for $t > t_o$, as well as for $t < t_o$. Note that within the context of our biological problem $0 < x_o < K$, but that the solution function (1.5,5d) is also defined for $x_o > K$ and for $x_o < 0$: draw a graph of our function for such values of x_o as well as for $x_o \in]0,K[$.

Question 1.5,7. One way of looking at a differential equation like (1.5,5c) is the following: take any point in the (t-x)-plane; then equation (1.5,5c) allows you to calculate the slope, $\dot{x}(t)$, of the solution curve which is supposed to pass through that point. Suppose this is done (do it!) at a great number of points in the (t-x)-plane, and suppose one draws a short line segment through each point in order to represent the corresponding slope at that point. In this way one obtains a picture strongly reminiscent of the old iron filings experiments in



elementary physics. Execute all this and compare the results with the graphs obtained in answer to Question 1.5,6.

Question 1.5,8. Compare two solution curves with the same γ, K and x_o, but different t_o-values. How do the two curves differ?

Discussion Question 1.5,9. The specific choice, $K-x$, for the function $f(x)$ of equation (1.5,5b) was made above for the somewhat flimsy reason that $K-x$ is simple. Several rationalizations have been proposed in the past, none of them very convincing besides a reasonably good fit to data (e.g., see Pearl and Reed, 1920, quoted in Gause's little book). An old (around 1839) argument by Verhulst postulated that in an agricultural society the relative rate of increase would be uninhibited until new farmland for farmers' sons would cease to be available--which would happen at a critical population size, say b; from that moment on, the relative rate of increase would be diminished by an amount directly proportional to the excess of actual over critical population size. Show that this postulate leads to an expression for $f(x)$ of the form $K-x$, or does it?

Discussion Question 1.5,10. Whatever the processes are that cause a population's increase to slow down, it seems a pretty safe bet to assume that their effect is not instantaneous. Accordingly, it has been proposed to replace the logistic equation by

$$\dot{x}(t) = \gamma\, x(t)\, (K - x(t-\tau)), \quad \text{for some } \tau > 0.$$

This would express, albeit in a starkly simplified way, that the decrease of the relative rate of increase is governed by the population size in the past. Discuss how the presence of the term $x(t-\tau)$ compels you to modify the mathematical approach of Question 1.5,7. You should find that now you cannot draw a direction field like the one in Question 1.5,7, but you have to construct every individual solution step by step, using its values at t and at $t-\tau$. How about the initial conditions? How would the solutions of the logistic equation compare with the solutions of the above modified form for identical values of γ and K?

1.6. Discrete-time vs. continuous-time models.

In Remark 1.5,3 we alluded to the existence of problems around the comparison of discrete-time and continuous-time models. These problems arise at two different levels, one of computing and the other of model building. Mathematically speaking, the two are related, though not identical. An important difference between the two is the role of the step length. Discretization for computational purposes is caused by the fact that digital computers can not represent continuous variables, but there is no conceptual lower bound to the size of the step length (although there is a practical lower bound), and so in many (though not all) cases problems arising from this discretization can be conquered satisfactorily by making the step length small enough (in the sense that the resulting approximation is then 'good enough'). Discretization for modeling purposes (as was indicated in Remark 1.5,3) has its origin in the biologists' feeling that the model should exhibit the same discreteness as the object of their studies. Thus the step length is determined by the length of the period between observations, or by the generation length, or by similar non-negotiable quantities. Thus the biologists will object to a smaller (or larger) step length than is 'natural' in the context of their work. Therefore, the problem arising from this kind of discretization cannot in general be attacked through diminishing the step length. We will list a few questions to enable the audience to discuss some aspects of these difficulties; for more information see H. R. van der Vaart (1973). A student project could easily grow from this topic. Some related questions are discussed in May (1973b).

Question 1.6,1. The differential equation

$$\dot{x}(t) = \lambda\, x(t), \quad x(0) = \xi,$$

has the solution

$$x(t) = e^{\lambda t} \cdot \xi.$$

The difference equation

$$x(t+h) - x(t) = h\, \lambda\, x(t), \quad x(0) = \xi\,,$$

from which the above differential equation arises as $h \downarrow 0$, has the solution

$$x(nh) = (1 + h\lambda)^n \cdot \xi\,,$$

which for $\lambda > 0$, $h > 0$ behaves as $e^{\lambda t}\cdot\xi$, but for $\lambda < 0$, $h > 0$ behaves very badly as soon as $|\lambda h| > 1$, and even worse as $|\lambda h| > 2$. Corroborate these statements.

Question 1.6,2. The difference equation

$$x(t+h) - x(t) = (e^{\lambda h} - 1)\cdot x(t), \quad x(0) = \xi ,$$

as $h \downarrow 0$, yields the same differential equation that was discussed in the preceding question. Show that the solution of this difference equation behaves exactly like the solution of $\dot{x} = \lambda x$, no matter what values λ and h have.

Remark 1.6,1. The preceding two questions illustrate the general principle that more than one (in fact, many) difference equations exist, all of which will yield the same differential equation as $h \downarrow 0$, but whose solutions may behave quite differently from those of that differential equation, when the step length h is finite. The problem for the discrete-model builder is how to select the 'right' difference equation.

Question 1.6,3. Lest the audience should think that decreasing the step length will always improve the degree to which the solutions of a difference equation approximate the solutions of the corresponding differential equation, a study should be made of the following difference equation:

$$x(t+2h) - 4\,x(t+h) + (3-2\alpha h)x(t) = 0,$$

or equivalently

$$\frac{1}{2h}[x(t+2h) - 4\,x(t+h) + 3\,x(t)] = \alpha\,x(t)$$

a) Show that if $x(\ )$ is differentiable, then the first member has the limit $-\dot{x}(t)$ as h tends to zero. (Hint: use truncated Taylor series with remainder term $o(h)$). Thus this difference equation seems to be a candidate for generating approximate solutions for the differential equation $\dot{x}(t) + \alpha\,x(t) = 0$.

b) Show that the condition $x(0) = \xi$ is not sufficient as initial condition, but you need the value of $x(\ )$ at one more value of its argument, e.g. $x(h) = \eta$ (now your calculator can find $x(2h)$, then $x(3h)$, etc.). Of course, if we want the solution to our difference equation to approximate the solution to $\dot{x}(t) + \alpha\,x(t) = 0$, $x(0) = \xi$, then we had better be careful in choosing the value η for $x(h)$. Perhaps one might suggest $x(h) = \xi e^h$.

c) Show that every solution to our difference equation is of the form,

$$c_1\langle 2 - \sqrt{1 + 2\alpha h}\rangle^{t/h} + c_2\langle 2 + \sqrt{1 + 2\alpha h}\rangle^{t/h}$$

for $t = 0, h, 2h, 3h, \ldots$; using our difference equation as a recurrence relation, we'll find the values of $x(\ )$ at only these values of t.

d) As $h \downarrow 0$, t fixed, the power in the first term has a limit, viz., $e^{-\alpha t}$ (prove!), which is encouraging (why?); but the power in the second term does not have a limit for fixed t: it behaves asymptotically (for fixed t; $h \downarrow 0$) as

$$3^{t/h}\, e^{\alpha t/3}$$

the quotient $\langle 2 + \sqrt{1 + \alpha h}\rangle^{t/h}/3^{t/h}$ has the limit $e^{\alpha t/3}$ as $h \downarrow 0$ for fixed t; prove!). Now this is the point made at the beginning of this question. Indeed, as $h \downarrow 0$ for fixed t, $3^{t/h}$ explodes; so things get *worse* in this case as the *step length* h *decreases*.

e) The only way to avoid this catastrophe is to keep $c_2 = 0$, but in the practice of numerical analysis this is impossible. Replacing a differential equation by a difference equation is done because digital computers can handle recurrence relations very well. In other words, in numerical work the solutions $c_1(\ldots)^{t/h} + c_2(\ldots)^{t/h}$ will not be used, but the computer will be given the values $x(0) = \xi$, $x(h) = \eta$, and then churn away. Now comes the clinch: even if η were 'equal' to $\xi(2 - \sqrt{1 + 2\alpha h})$, still the computer, working with a finite word length (resulting in rounding-off error), will after some steps obtain values that do slightly differ from the 'ideal' value $\xi\langle 2 - \sqrt{1 + 2\alpha h}\rangle^{t/h}$; then this difference, however small, will *explode* according to the factor $3^{t/h}$, *faster* as *h* is *smaller*.

f) The problem sketched in this question is part of the topic 'Numerical stability of linear multistep methods'; e.g., see section 3.2 in L. *Lapidus* and J. H. *Seinfeld* (1971).

g) Let us reiterate once more: the role of this discussion in the present context is to show that even when we can manipulate the step length in a difference equation that is supposed to approximate a certain differential equation, the approximation does not necessarily become better, in fact, it may get worse.

Now we will return to our original point: a difference equation which yields a certain differential equation when, after the fashion of all physical science models, its step length h goes to zero, may have solutions that for finite h have a wildly different behavior from the solutions of the ('limiting') differential equation.

Question 1.6,4. The differential equation

$$\dot{x}(t) = -[x(t)]^2, \quad x(t_o) = \xi,$$

has the solution

$$x(t) = \frac{1}{\frac{1}{\xi} + t - t_o};$$

explore its behavior for $t > t_o$ as $\xi = 200$, as $\xi > 0$, as $\xi = 0$, and as $\xi < 0$.

The difference equation

$$x(t+h) - x(t) = -h\,[x(t)]^2, \quad x(t_o) = 200, \quad h = 0.01;$$

can be written as

$$x(t+h) = x(t)\,[\quad ? \quad].$$

Compute $x(t_o + .01)$, $x(t_o + .02)$, $x(t_o + .03)$, and $x(t_o + .04)$, and compare your results with the solutions of the differential equation.

Do the same with the difference equation,

$$\frac{1}{x(t+h)} - \frac{1}{x(t)} = h, \quad x(t_o) = 200, \quad h = 0.01.$$

Also show that both of these difference equations yield the differential equation $\dot{x} = -x^2$ when $h \downarrow 0$.

Question 1.6,5. Show that the solution (see Question 1.5,4) of the logistic equation

$$\dot{x}(t) = \rho\, x(t) \left\langle 1 - \frac{x(t)}{\alpha} \right\rangle, \quad x(0) = \xi \in \,]0,\alpha[,$$

coincides at the points $t = 0, h, 2h, 3h,$ etc. with the solution of

$$\frac{\alpha}{x(t+h)} - 1 = \left\langle \frac{\alpha}{x(t)} - 1 \right\rangle e^{-\rho h}, \quad x(0) = \xi \in \,]0,\alpha[.$$

Also show that the latter equation can be written as

$$x(t+h) = \frac{\lambda\, x(t)}{1 + \beta\, x(t)}, \quad \text{with} \quad \lambda = e^{\rho h}, \quad \beta = ?$$

Discussion Question 1.6,6. The last equation of the preceding question has been used by Leslie (1958) as the point of departure for a stochastized version of the discrete logistic model. Given $x(\ )$ at time t, he determined $x(t+h)$ as a value of a normal random variable with mean $\lambda\, x(t)\cdot[1 + \beta\, x(t)]^{-1}$ and variously chosen variances, all of which he took to be proportional to the mean of $x(t+h)$ (this proportionality is strictly speaking a property of linear birth and death processes, which Leslie apparently extrapolated to the present logistic process). With a table of random numbers and a calculator, it will be easy to experiment with different-sized variances and compare the result with the corresponding deterministic logistic.

2. PREDATOR-PREY

2.1. Lotka-Volterra model.

Consider an ecosystem consisting of two species of animals, one of which (the prey) is eaten by the other (the predator). Let x represent the number of prey and y, the number of predators. Suppose for a moment that there are no predators in the system, just individuals of the prey species, with plenty of food and space. Then one could imagine that the prey population would increase according to the Malthusian (= geometric-logarithmic-exponential) pattern:

$$\dot{x}(t) = x(t)\cdot\gamma_1 \tag{2.1,1a}$$

where γ_1 would be a positive constant, the relative rate of increase. Now suppose there are also some predators in the system (the case we are interested in). Then there will be encounters between individuals of the prey species and individuals of the predator species. A certain fraction of these encounters will have fatal results for the prey individuals. So $\dot{x}(t)$, the rate of increase of the prey population, must be less than $x(t)\cdot\gamma_1$, and the difference should be proportional to the number of encounters per time unit. Now it stands to reason that the number of such encounters during a short time interval starting at time t is a monotone increasing function of $x(t)$

itself (i.e., the number of prey individuals at time t), as well as of $y(t)$ (i.e., the number of predator individuals at time t). The only unresolved problem is then the choice of a function of x and y, monotone in x and monotone in y, to represent the number of encounters per time unit. In analogy to chemical kinetics (e.g., see Boudart, 1968, pp. 42 and 51), Lotka (1956, p.88) assume this function to be $x \cdot y$. Thus equation (2.1,1a) has to be amended and replaced by

$$\dot{x}(t) = x(t) \cdot [+ \gamma_1 - p\, y(t)]. \qquad (2.1,1b)$$

So much for the prey. Now suppose for a moment that there are no prey animals in our system, just predators. These predator individuals, being without food, would starve. So their population would decline. Let us assume, somewhat rashly, that their decline would obey this description:

$$\dot{y}(t) = y(t) \cdot (- \gamma_2), \qquad (2.1,1c)$$

where γ_2 would be a positive constant. Now suppose there are also some prey individuals in the system. Then the predators would be doing better. In fact, they would be doing better in proportion to the number of encounters with prey animals. So the value of $\dot{y}(t)$ should be increased by an amount proportional to the number of encounters between predator and prey. According to Lotka's assumption this means that equation (2.1,1c) should be amended and replaced by

$$\dot{y}(t) = y(t) \cdot [- \gamma_2 + q\, x(t)] \qquad . \qquad (2.1,1d)$$

Remark 2.1,1. Volterra (1927, p. 38) does not adduce chemical kinetics in defense of the terms $+ q\, x\, y$ or $- p\, x\, y$. He does remark that the relative rate of increase should be less than $+ \gamma_1$ for the prey (because of the predators) and greater than $- \gamma_2$ for the predator (because of the prey); then he goes on to say that the simplest (albeit rather crude) way to achieve this is by subtracting a term proportional to $y(t)$, and adding a term proportional to $x(t)$, respectively.

Remark 2.1,2. One could also, and perhaps more impressively, argue that a population of $y(t)$ predator individuals needs a total feeding rate of $y(t) \cdot \gamma_2$, just to maintain itself at that level. So it is only the excess over that amount that contributes to the growth of the predator population:

$$\dot{y}(t) = q\, x(t)\, y(t) - \gamma_2 y(t),$$

the same equation as (2.1,1d).

Remark 2.1,3. Equation (2.1,1a), representing the growth of the prey population in the absence of predators, lacks realism in that no prey population would keep growing exponentially. So we really need an expression that will not exhibit unlimited growth in the absence of a predator (or even in the presence of a predator). The traditional tool to represent a limited growth is the logistic; for a prey population without predators:

$$\dot{x}(t) = x(t)\cdot[+\gamma_1 - \delta\, x(t)]. \qquad (2.1,1e)$$

The argument that led from equation (2.1,1a) to equation (2.1,1b), will now lead to

$$\dot{x}(t) = x(t)\cdot[+\gamma_1 - \delta\, x(t) - p\, y(t)],$$

or in short notation,

$$\dot{x} = x\cdot(+\gamma_1 - \delta\, x - p\, y). \qquad (2.1,1f)$$

A similar argument might suggest that equation (2.1,1d) be replaced by

$$\dot{y} = y\cdot(-\gamma_2 + q\, x - \zeta\, y) \qquad (2.1,1g)$$

Question 2.1,1. The "Volterra prey-predator equations" are (2.1,1b) and (2.1,1d) together,

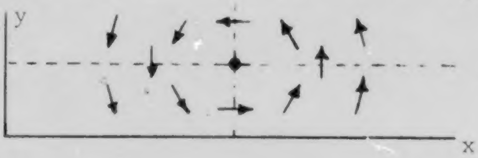
$$\begin{cases} \dot{x} = x\cdot(+\gamma_1 \qquad\quad -p\, y) \\ \dot{y} = y\cdot(-\gamma_2 + q\, x \qquad\quad) \end{cases}, \qquad (2.1,2)$$

with $\gamma_1 > 0$, $\gamma_2 > 0$, $p > 0$, $q > 0$. We want to study some of the properties of the solutions of this system. In the first place show that a scale change of x and y ($x = \beta_1 x_1$, $y = \beta_2 x_2$) makes it possible to write (2.1,2) in the form

$$\begin{cases} \dot{x}_1 = x_1\cdot(+\gamma_1 \qquad\quad + \alpha_{12} x_2) \\ \dot{x}_2 = x_2\cdot(-\gamma_2 + \alpha_{21} x_1 \qquad\quad) \end{cases} \qquad (2.1,3)$$

with γ_1 and γ_2 same as before, and $\alpha_{21} = \alpha > 0$, $\alpha_{12} = -\alpha_{21} = -\alpha$. The same kind of transformation will play a role later in models for n species systems.

Question 2.1,2. Apply the phase plane-isocline-arrow method (see appendix) to equation (2.1,2) and see how it leads you to expect a more or less cyclic behavior.



Question 2.1,3. Equations (2.1,2) can be treated analytically, not to express x and y as functions of t, but to find a relation between x and y, which in turn determines a family of closed curves in the (x-y)-plane. In fact, show that equations (2.1,2) for all real values, □, of t entail

$$\frac{\dot{x}(\square)[-\gamma_2 + q\ x(\square)]}{x(\square)} = \frac{\dot{y}(\square)[+\gamma_1 - p\ y(\square)]}{y(\square)}$$

Show that integration of this equality from t_o to t yields

$$\left(\frac{x(t)}{x_o}\right)^{-\gamma_2} \cdot \frac{e^{q\ x(t)}}{e^{q\ x_o}} = \left(\frac{y(t)}{y_o}\right)^{+\gamma_1} \cdot \frac{e^{-p\ y(t)}}{e^{-p\ y_o}} \qquad (2.1,4)$$

Show that for given (x_o, y_o) this equation determines a closed curve* e.g., see p. 84 of d'Ancona, 1954). Also make clear that with varying (x_o, y_o) this equation generates a family of cycles (closed curves), filling up the entire first quadrant. Thus the analytical work (which in this case was possible) confirms the quick-and-dirty graphical analysis mentioned in Question 2.1,2; as of course it should.

Remark 2.1,4. The audience has just become acquainted with a basic fact of life in mathematics as applied to the life sciences: most of the differential equations that one encounters there cannot be solved explicitly; most not even implicitly. This is frustrating to some people, but it is very important for anybody going into applied mathematics to come to terms with it. After all, that is why computers were invented and built! One more thing: at first sight one might feel that mathematics as applied to physics has not had this problem; but this is only in appearance. In the classical period of mathematical physics each new differential equation that arrived on the scene was greeted by the creation, or rather canonization, of a new function (e.g., the functions of Mathieu, Lamé, Bessel, etc.); if you give a name (and a symbol) to a typical solution of the canonic form of a (linear) differential equation that you could not solve explicitly before that name and that symbol were given, then you can solve it 'explicitly' afterward. So even the physicists in their classical period knew this basic problem. After the nonlinear equations came up (which happened very early in the development of mathematical biology, whereas it took a long time in mathematical physics), even the physicists stopped their name game; so some of their research is now in the same situation as our audience finds itself in.

---

* The proof is fairly long, although elementary. It is in the quoted location of U. d'Ancona (1954). This book was listed as essential. So there is little point in our reproducing the proof here.

Remark 2.1,5. After the mathematical community realized that there would never be an end to the list of 'special functions of mathematical physics,' more emphasis was put on determining certain properties of the collection of all solutions of a given (set of) differential equation(s) without first determining any kind of analytical expression for the solutions; viz., such properties as number and nature of singular points, existence of periodic solutions, boundaries of regions of attraction of stable singular points, etc., which properties are variously called geometric or qualitative properties; more rigorously, properties that are preserved under arbitrary topological transformations of the phase space. Since the collection of all trajectories in the phase space (i.e., (x,y)-space in the above example) is called the phase portrait of the differential system, we may summarize the situation by saying that as the importance of nonlinear systems grew, the study of topological properties of their phase portraits grew right along with it. This is not to say that numerical solutions obtained by digital or analog computer are no longer wanted or useful, or even that analytical methods should no longer play a role, but it does say that nobody should feel frustrated if mathematical models in ecology, or in any part of biology, or in any other field, are not solvable in closed form; the mathematical community has shed that frustration some time ago. A large number of textbooks at various levels have appeared, dealing with such questions, e.g., Brauer and Nohel, 1969 (aimed at undergraduates); Hurewicz, 1958 (paperback, slightly more advanced); Aggarwal, 1972 (paperback, aimed at seniors); Andronov, Leontovich, Gordon, Maier, 1973 (expensive, advanced, excellent).

Question 2.1,4. The amended Volterra prey-predator equations (including 'logistic terms') are (2.1,1f) and (2.1,1g) together,

$$\left\{\begin{aligned} \dot{x} &= x\cdot(+\gamma_1 - \delta x - p y) \\ \dot{y} &= y\cdot(-\gamma_2 + q x - \zeta y) \end{aligned}\right\}, \qquad (2.1,5)$$

where $\gamma_1, \gamma_2, \delta, \zeta, p, q$ represent positive numbers. In Question 2.1,1 you showed that a scale change in x and y could transform the matrix of coefficients of x (or x_1) and y (or x_2) of equation (2.1,2) into a skew-symmetric matrix,

$$\begin{pmatrix} 0 & \alpha_{12} \\ \alpha_{21} & 0 \end{pmatrix}, \quad \text{with } \alpha_{12} = -\alpha_{21}.$$

This is entirely impossible for equation (2.1,5). Why?

Question 2.1,5. Apply the phase plane-isocline-arrow method (see appendix) to equation (2.1,5), distinguishing two main cases,

$$\frac{\gamma_2}{q} < \frac{\gamma_1}{\delta}, \quad \text{and} \quad \frac{\gamma_2}{q} > \frac{\gamma_1}{\delta},$$

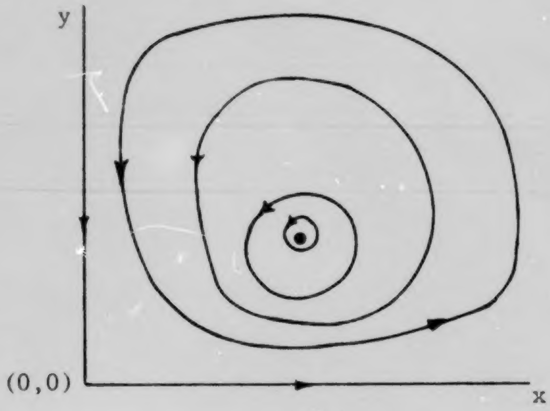
and the limiting case in between: $\frac{\gamma_2}{q} = \frac{\gamma_1}{\delta}$.

Question 2.1,6. Follow the work you did in Question 2.1,2 and in Question 2.1,5 by the study of the matrix of the linear systems that approximate the given nonlinear systems in the neighborhood of each of their singular points (see section A.4 of Appendix A). Do you find any contradictions?

Discussion Question 2.1,7. The reader should explore why the integration method(s) discussed in Question 2.1,3 (and in Question 2.1,10 below) do not help in the case of equation (2.1,5). This does not, of course, mean that no solution curves $\langle x(t), y(t) \rangle$ exist, but that they (or rather, their projections onto the (x,y)-plane) cannot be represented by simple analytical expressions.

Discussion Question 2.1,8. In Question 2.1,5 you found that the two main cases exhibit a major difference in phase portrait, a major difference also in biological outcome. If you'd want to tell a biologists about this result, he would react with the question: now what do these inequalities (viz. $\frac{\delta}{\gamma_1} \gtrless \frac{q}{\gamma_2}$) mean biologically? What would your answer be?

Discussion Question 2.1,9. In Hirsch and Smale (1974), an excellent text on differential equations, the following figure is given as phase portrait of equation (2.1,2):



Any comments?

Question 2.1,10. A different way of finding a result equivalent to the one found in Question 2.1,3 takes off from the transformed equations (2.1,3) in Question 2.1,1. First determine (ξ_1, ξ_2) as the solution of

$$\left\{\begin{array}{lllr} +\gamma_1 & & +\alpha_{12}x_2 & = 0 \\ -\gamma_2 & +\alpha_{21}x_1 & & = 0 \end{array}\right\}.$$

Since (see Question 2.1,1) $\gamma_1 > 0$, $\gamma_2 > 0$, $\alpha_{21} > 0$, and $\alpha_{12} < 0$, it follows that both ξ_1 and ξ_2 are positive. Then show that equations (2.1,3) can be rewritten in the form

$$\left\{\begin{array}{l} \dot{x}_1/x_1 = \alpha_{12}(x_2 - \xi_2) \\ \dot{x}_2/x_2 = \alpha_{21}(x_1 - \xi_1) \end{array}\right\}. \tag{2.1,6}$$

In equations (2.1,6) make the substitution $x_i = \xi_i e^{v_i}$ $(i = 1,2)$; premultiply the first of the resulting equations by $(e^{v_1} - 1)\xi_1$ and the second by $(e^{v_2} - 1)\xi_2$; add the two and find

$$\left(e^{v_1(t)} - 1\right) \xi_1 \dot{v}_1(t) + \left(e^{v_2(t)} - 1\right) \xi_2 \dot{v}_2(t) = 0 \tag{2.1,6a}$$

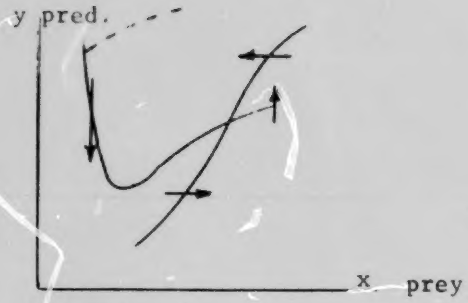
Upon integration find

$$\left[e^{v_1(t)} - v_1(t)\right] \xi_1 + \left[e^{v_2(t)} - v_2(t)\right] \xi_2 = \text{constant} \tag{2.1,6b}$$

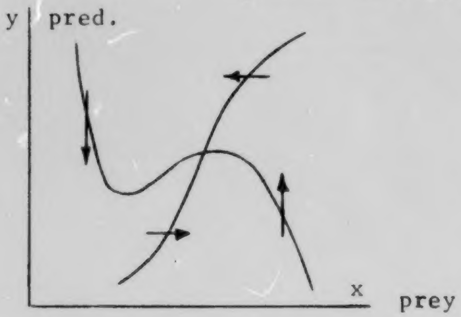
Remark 2.1,6. The first member of equation (2.1,6b) is a 'constant of motion' or a 'first integral.' This result is essentially equivalent to equation (2.1,4) (see Question 2.1,3), although the present method is easier to generalize to systems of more species, and this result also shows (as did equation (2.1,4)) that the systems defined by equations (2.1,2) and (2.1,3) are 'conservative.' This has excited some people with a physical background very much (Kerner, 1972; Goel, Maitra and Montroll, 1971), because they saw an opportunity here to apply methods of classical and statistical mechanics to these ecosystem models. Unfortunately, models that are even slightly more realistic (such as equation (2.1,5), see Question 2.1,4 and Remark 2.1,3) are no longer conservative, so that at present it is doubtful exactly how useful these attempts are. Equally unfortunately, nobody has managed to find any biological meaning whatever for the 'constant of motion' of equation (2.1,6b): Kerner himself conceded this during a conference on theoretical biology (cf. p. 155 of Jacobs, 1966).

2.2. Critique of the Lotka-Volterra Model, and proposed improvements.

Experiments performed (a.o., by Gause 1934, 1971) in order to check the theoretical prediction of oscillations, have shown that in the case of homogeneous environments (such as easily occur in laboratory experiments) the predator species will often succeed in finishing off the prey species, contrary to the theory. But if an environment is made complex enough (or better: inhomogeneous enough) so that the prey species has a hiding place (refuge) where a certain number of prey may retreat and be unattainable for the predator species, at least temporarily, then oscillations do occur (for instance see Huffaker, 1958, or Flanders and Badgley, 1963). This caused MacArthur and Connell (1966) to come up with isoclines of the type discussed in Questions 2.2,1 and 2.2,2.



Question 2.2,1. The reader may show that that portion of the 'vertical' isocline which is pointed out by the broken arrow, does make it impossible for the predator to run the prey into the ground.



Question 2.2,2. According to further specification, the 'vertical' isocline would, moreover, bend its rightmost end downwards (see accompanying diagram). What biological realism, if any, would be added by this bend?

Kolmogorov (1972) did not so much criticize the Lotka-Volterra model as propose a generalization of it, in a direction somewhat different from the above equation (2.1,5). Again with x for the number of prey and y for the number of predators, he proposes

$$\left\{ \begin{aligned} \dot{x} &= K_1(x)\cdot x - L(x)\cdot y \\ \dot{y} &= K_2(x)\cdot y \end{aligned} \right\} \qquad (2.2,1)$$

Question 2.2,3. Show that the Lotka-Volterra model (equation (2.1,3)) is indeed a special case of this equation (2.2,1). Also show that the amended model which included 'logistic' crowding terms (see equation (2.1,5)) is a special case of equation (2.2,1) except for one term. Show how the above-given motivation for this one term indeed clashes with Kolmogorov's stated assumptions underlying his model as given by equation (2.2,1); indeed, Kolmogorov (l.c.) based his equation (2.2,1) on the following assumptions:

(1) The predators act independently of each other; therefore their relative rate of increase, K_2, and the rate at which one predator kills prey, L, are constant functions of y, though not of x.

(2) The rate of increase ($\dot{x}$) of the prey in the presence of the predators equals their rate of increase in the absence of the predators

minus the rate at which the prey population loses members because of the killing by the predators.

Discussion Question 2.2,4. The 'horizontal' isocline (i.e., the locus of points where the trajectories have horizontal tangents) is given by

$$K_2(x) \cdot y = 0.$$

How does this equation compare with Kolmogorov's sketches of possible phase portraits in the attached photocopy? ($N_1 = x$; $N_2 = y$).

Discussion Question 2.2,5. The 'vertical' isocline (i.e., the locus of points where the trajectories have vertical tangents) is given by

$$K_1(x) \cdot x - L(x) \cdot y = 0.$$

Whether or not the y-axis belongs to this isocline, depends on the function $L(\ )$; how? The case that the y-axis is not a part of the vertical isocline provides an example of Remark 1.5,4. (For another example see the diagrams in Holling and Ewing (1971), which are the result of machine simulation, not of a model such as equation (1.5,3).

Discussion Question 2.2,6. Before it is possible to discuss if the sketches of possible phase portraits in the attached photocopy are compatible with the equation of the 'vertical isocline', we had better know a little more about the functions $K_2(\ )$, $K_1(\ )$, and $L(\ )$. Kolmogorov makes the following assumptions regarding them. Discuss if they seem to be reasonable.

$K_1(x)$ a monotone decreasing function of x; $K_1(0) > 0 > K_1(+\infty) > -\infty$.

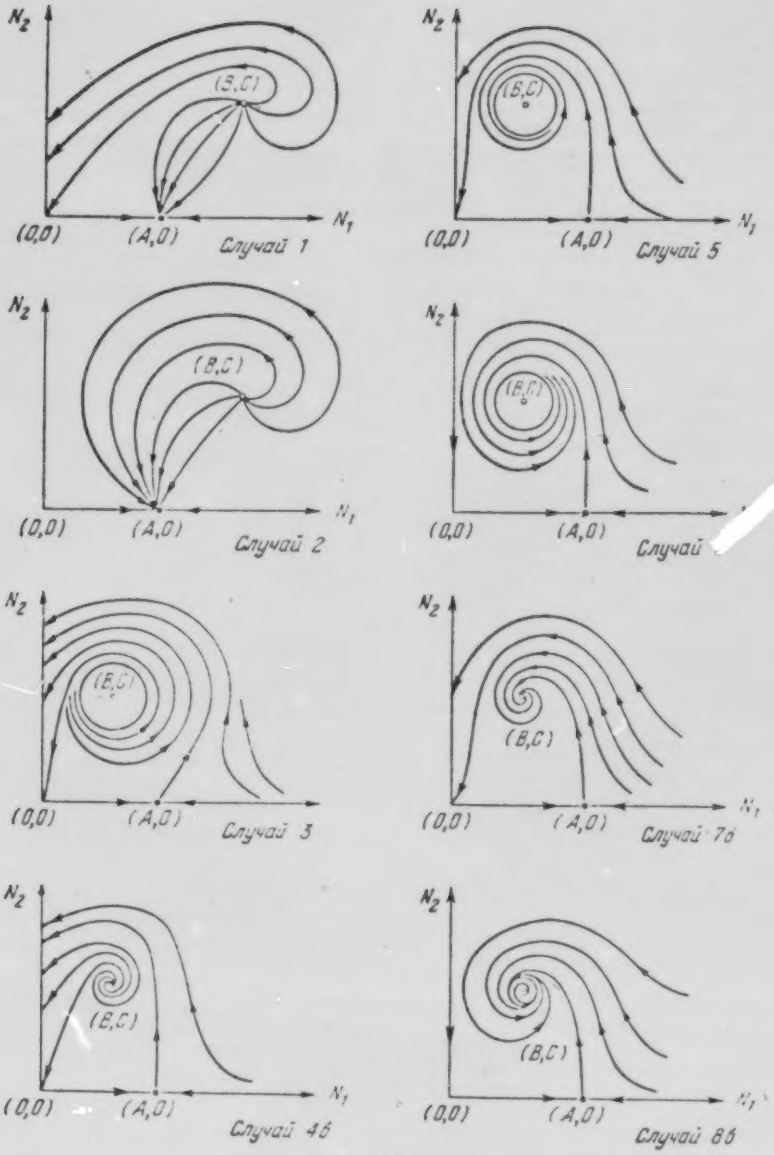
$K_2(x)$ a monotone increasing function of x; $K_2(0) < 0 < K_2(+\infty)$.

$L(x) > 0$ for $x > 0$; $L(0) = 0$ or $L(0) > 0$.

Now in the attached photocopy the cases numbered 2, 6, 8 are supposed to represent the case $L(0) = 0$; the others, $L(0) > 0$. Now you can check if the sketches do well by the 'vertical' isocline.

You can also check the location of the singular points, as well as their nature (by the linearization method). For the saddle points among them you can find the two asymptotic directions of trajectories passing through them.

Finally in the cases 1, 3, 4, 5, 7 check the angles under which the trajectories hit the y-axis.



N2
N1
(B,C)
(0,0)
(A,0)
Случай 1
N2
N1
(B,C)
(0,0)
(A,0)
Случай 5
N2
N1
(B,C)
(0,0)
(A,0)
Случай 2
N2
(B,C)
(0,0)
(A,0)
Случай
N2
(B,C)
(0,0)
(A,0)
Случай 3
N2
N1
(B,C)
(0,0)
(A,0)
Случай 7б
N2
N1
(B,C)
(0,0)
(A,0)
Случай 4б
N2
N1
(B,C)
(0,0)
(A,0)
Случай 8б

Photocopy of p. 105 in Kolmogorov (1972).

His N_1 is our x ; his N_2 is our y .

Re this last point, Kolmogorov notes that as the equations are written the trajectories would go on to negative x-values, which is biologically impossible. Therefore, he writes, one should interpret the equations in this way that once the trajectory has hit the y-axis, the first equation stops action and from there on in only the second equation acts, for $x = \text{constant} = 0$.

Leslie (1958, p.20) objects to equation (2.1,2) on grounds related to those named in Remark 2.1,3: No allowance is made for intra-specific competition (which is easily remedied by inserting negative terms, $-x^2$ and $-y^2$) and, more seriously, there is "no upper limit to the relative rate of increase of the predator" (which objection still holds against the amended Lotka-Volterra equations (2.1,5), although for any given x there is such an upper limit; but then again, x itself is variable). So Leslie wants the relative rate of increase of the predator to equal a constant diminished by a certain term; this term, since the predator species does worse as the ratio of predator individuals to prey individuals is larger, Leslie proposes to choose proportional to $\frac{x}{y}$. Therefore his equation is

$$\left\{\begin{array}{ll} \dot{x} = x\cdot(r_1 - a_1 x - b_1 y) & \text{for the prey} \\ \dot{y} = y\cdot\left(r_2 - a_2 \frac{x}{y}\right) & \text{for the predator} \end{array}\right\}. \qquad (2.2,2)$$

Question 2.2,7. Sketch the phase portrait of this equation. Is it biologically more acceptable than the phase portrait of equation (2.1,5)? What happens at the points of the x-axis? What are the singular points? Execute the linearization analysis in their neighborhoods.

Discussion Question 2.2,8. If you find the phase portrait of equation (2.2,2) to be less acceptable than that of equation (2.1,5), can you then pinpoint a flaw in the above-reproduced argument of Leslie?

Discussion Question 2.2,9. Leslie (1958) also touches on the topic we mentioned in Remark 1.5,3 and in section 1.6: the preference that many biologists have for discrete over continuous models. An additional reason for Leslie's interest in discrete models is the fact that he wants to represent numerically random agents influencing the ecosystem along the lines indicated in Discussion Question 1.6,6. So how does Leslie choose a discrete model for a prey-predator system? In analogy to the discretization of the logistic introduced in Question 1.6,5. Just as

$$\dot{x}(t) = \rho\, x(t) \left(1 - \frac{x(t)}{\alpha}\right)$$

was (rigorously) replaced by

$$x(t+h) = \frac{\lambda\, x(t)}{1 + \beta\, x(t)} ,$$

so the first of the equations in equation (2.2,2) is (approximately) replaced by

$$x(t+h) = \frac{\lambda_1 x(t)}{1 + \alpha_1 x(t) + \beta_1 y(t)} \qquad (2.2,3a)$$

and the second of the two is replaced by

$$y(t+h) = \frac{\lambda_2 y(t)}{1 + \alpha_2 [x(t)/y(t)]} \qquad (2.2,3b)$$

The role of the isoclines is now played by the locus of points (ξ,η) in the phase plane such that $x(t) = \xi$ implies that $x(t+1) = \xi$ ('vertical' isocline), or that $y(t) = \eta$ implies that $y(t+1) = \eta$ ('horizontal' isocline). Compare the isoclines, thus defined, for equations (2.2,3 a and b) with the isoclines (in the usual sense) for equations (2.2,2).

Another obvious reason for criticizing any and all of the predator-prey models discussed so far is the fact that the food of the prey is not represented. This may be alright if the food of the prey is plentiful, but it gives one a lack of trust in the realism of these models, especially if the prey itself is a predator of some other species. Also there is an oddity in Volterra's modeling of a food chain: it has strange difficulties if the chain consists of an even number of species. These difficulties disappear for the larger part if the food of the prey is also represented in the model, even if the food consists of plants. Poletaev (1966) and Eman (1966) are among the authors that have explicitly addressed themselves to these problems. Eman proposed several models that (at the same level of crudeness as the original Volterra models) took explicit account of the environment of the predator and prey species: e.g., plants and soil. For a discussion of one or two models of this type see Chapter 4.

Looking back to the first few pages of section 2.1, we see clearly that the arguments for the 'interaction' terms being proportional to $x \cdot y$ were not strong: an analogy to chemical kinetics, a remark that this choice is simple (though crude). No wonder that Kolmogorov suggested the qualitative investigation of more general models. On the other hand, one may, by changing these interaction terms, achieve the explicit solvability of the equations. For instance, Gomatam (1974a) published a paper on the model (cf. equation (2.1,2)):

$$\left\{ \begin{aligned} \dot{x} &= + \gamma_1 x - b_1 x \log y \\ \dot{y} &= - \gamma_2 y + b_2 y \log x \end{aligned} \right\}, \tag{2.2,4}$$

and on the model (cf. equation (2.1,5)):

$$\left\{ \begin{aligned} \dot{x} &= x(+ \gamma_1 - b_{11} \log x - b_{12} \log y) \\ \dot{y} &= y(- \gamma_2 + b_{21} \log x - b_{22} \log y) \end{aligned} \right\} \tag{2.2,5}$$

Question 2.2,10. After the substitutions $\log x = \xi$ and $\log y = \eta$ these equations reduce to linear equations and can be solved explicitly. Do so and discuss the properties of the corresponding solutions for $x(t)$ and $y(t)$ as compared to the results obtained on the original Lotka-Volterra models of the equations (2.1,1) and (2.1,5).

None of these devices, however, are truly agreeable to a biologist for the simple reason that they do not reflect any biological knowledge or data concerning predator-prey interaction. Correspondingly Holling (1966), in a detailed paper, which also discusses some of the older attempts in this area, conducted an experimental study of the interaction of one particular predator species and one particular prey species. As an example of the approach to modeling that biologists really like, this would be a good paper to study. Unfortunately, its discussion here would clearly exceed the boundaries of the present writing project.

It seems possible and useful, though, to have a brief discussion of some points made by Watt (1959). (This paper is mainly on parasites rather than predators, but both parasites and predators have to find their victims before they can attack them, whether for the purpose of egg-depositing or eating; so for the discussion on interaction his remarks are relevant, as the reader can see for himself): the Lotka-Volterra assumption that the number of prey attacked per time unit is proportional to $x \cdot y$, the product of the prey density and the predator density, is implausible for at least two reasons. *First*, it implies that for given predator density y the number of prey attacked is directly proportional to the prey density x, no matter how many prey there are. This is unsatisfactory, if only because even the biggest glutton can eat only so much: any digestivie system can take only a certain amount of food per time unit. Also, as hunger decreases a predator will search and attack less vigorously, etc. *Second*, the Lotka-Volterra assumption postulates that for given prey density x the number of prey attacked is directly proportional to the predator density y, no matter how many predators there are; in other words, that the prey-searching and attacking behavior of the predator individuals is not modified by increasing predator density: the predators are assumed to act independently of one another, no matter how many of them are around. Watt (l.c.) quotes many experimental studies showing that this assumption does not hold in the case of parasites, and the same is bound to be true for predators.

Several alternatives have been discussed in the literature. Relatively easily accessible is May (1973a), available in paperback. Other references, which may or may not be accessible to the reader, are Holling (1966), Watt (1959), Watt (1968), Royama (1971), Problemy Kibernetiki, especially vol. 25 (1972); this list does not aim at completeness. The *first* objection listed in the preceding paragraph may be met by replacing the Lotka-Volterra cross-product $x \cdot y$ by $\hbar \cdot y \cdot (1 - e^{-\beta x})$ (used by Gause, Ivlev, and others; note that for small x this function $\sim x \cdot y$): this expression puts an upper bound to the number of encounters between any one predator and individuals of the prey species when the prey grows abundant. Of course, the replacement of $x \cdot y$ by $k \cdot x \cdot y/(b+x)$ would give similar service. The *second* of the above objections may be met similarly by replacing $x \cdot y$ by $c \cdot x \cdot (1 - e^{dy})$ (used by Thompson, Nicholson and Bailey, and others; what is the situation for small y-values?)

Discussion Question 2.2,11. The reader should construct some alternative predator-prey models with the above building blocks and investigate their properties by means of the methods used before (isocline-arrow and linearization at singular points).

Another way in which to make our models more life-like would be to account for inhomogeneities. The habitat of a real population is not homogeneous: for instance there will often be hideaways for the prey, places where the predator has difficulty moving about. The populations are not homogeneous: age differences, differences in hunger condition, genetic differences, differences between the sexes. Each of these possibilities have been considered by one or more authors. One will find references in the journals quoted at the end of section 1.3.

An obvious necessity is to discuss ecosystems consisting of more than two species: outside the laboratory 2-species ecosystems are virtually nonexistent. We should also emphasize that species may enter into other relationships besides that of predator and prey. Chapter 3 will briefly discuss a few of those other relationships. Chapter 4 will mention some of the problems in the study of more than two species. The last section of Chapter 2 will indicate a more qualitative approach to predator-prey theory than any of the others.

2.3. General remarks about the predator-prey model.

It does happen that one is required to come up with a mathematical model for a situation about the mechanism of which very little is known. As an example of what one might be able to accomplish in such situations let's try to draw conclusions regarding the functions $\varphi(x,y)$ and $\psi(x,y)$ (and/or their ratio) in the general equations (1.5,3).

$$\left\{ \begin{array}{l} \dot{x} = \varphi(x,y) \\ \\ \dot{y} = \psi(x,y) \end{array} \right\} \qquad (2.3,1)$$

merely from the very general observation that a small prey population cannot sustain as large a predator population as a large one and that a large predator population eats more prey than a small one.

So, again with x the number of prey and y the number of predators, let us look at the rightmost vertical line in Figure 2.3,1. This line represents a number of different predator-prey systems, all with a sizable, though not excessive, number of prey, but with a varying number of predators, ranging from a very few (down) to many (up). A system at point a is a paradise for prey (very few predators); so the number of prey will increase somewhat; but it

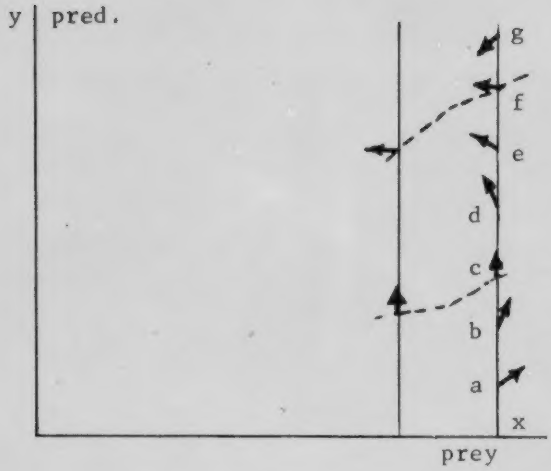


Figure 2.3,1

is maybe even more a paradise for the predator (abundant food supply): so the number of predators will increase as well, and the system will move as indicated by the arrow at a. The reader will be able to argue the other arrows, at the points b through f: as we look at systems with more and more predators, they become less and less of a paradise for the prey, which at c reach a situation where they can just hold their own against the inroads of the predator, and which beyond c will decrease: kill exceeds birth. As we look at systems with still more predators, there will come a point f, where the predators do not find enough food to increase: they can just hold the line. With still more predators present, the number of predators will even decrease: see point g.

Now look at a line a little bit to the left of the one we just discussed: less prey. So the point c, as well as the point f, will be reached earlier: fewer predators will now be able to prevent the prey population from increasing, and the predator population will fail to find enough food (since there is less of it) at a lower predator density.

> Question 2.3,1. What is the value of $\varphi(x,y)$ at any of the points c (where $\dot{x} = 0$), and of $\psi(x,y)$ at any of the points f (where $\dot{y} = 0$)? What can you say about the values of $\psi(x,y)$ at c and of $\varphi(x,y)$ at f?

Evidently the points of type c make up the 'vertical' isocline (points where $\dot{x} = 0$) and the points of type f make up the 'horizontal' isocline (points where $\dot{y} = 0$). The above argument suggests that in the right-hand part of the (x,y)-diagram the 'horizontal' isocline should be above the 'vertical' isocline and that both should tend to go to the left and down. This argument needs some supplementation in two ways.

In the *first* place, the above presentation was somewhat biased, or loaded. On the one hand, it is clear that in a collection of predator-prey systems, which have all a certain sizable, though not excessive, prey population, but different sizes of predator populations, the systems with a large predator population provide both species with a less attractive environment: the predators don't have enough to eat and the prey have a good chance of being killed. So as we consider different ecosystems with the same prey density, but with more and more predators, the rate of increase of the predator population will go down and the rate of increase of the prey population will go down; until both species wind up decreasing (viz., if the system has a large prey population as well as a large predator population). There is a question, however, as to which of the two species will be the first to hit zero rate of increase:

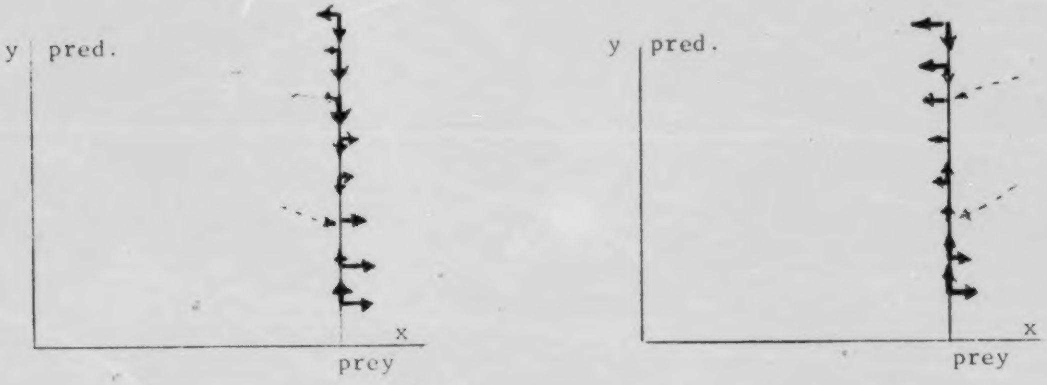


Figure 2.3,2; see text above and below

As is seen immediately, Figure 2.3,1 constitutes the second of these two possibilities (see the following remark).

Remark 2.3,1. The arrows have the same role as in the Appendix on the phase-plane isocline and arrow method: a horizontal arrow indicates increase or decrease of the prey population (x = prey); a vertical arrow indicates increase or decrease of the predator population (y = predator); their relative lengths represent the relative size of $\dot{x}$ and $\dot{y}$; and where the system is going is determined by the resultant. Now compare Figure 2.3,1 and the second diagram in Figure 2.3,2.

In the second place, we need to supplement the argument around Figure 2.3,1 with a second collection of predator-prey systems. All systems in this collection will have a fairly small predator population, but their prey populations will have different sizes. A system with few predators, but relatively many prey, will see both predators and prey increasing, but as we look at systems with fewer and fewer prey, both species will see their rate of increase go down, until both species wind up decreasing (viz., if the system has a small prey population as well as a small predator population). Again, two possibilities:

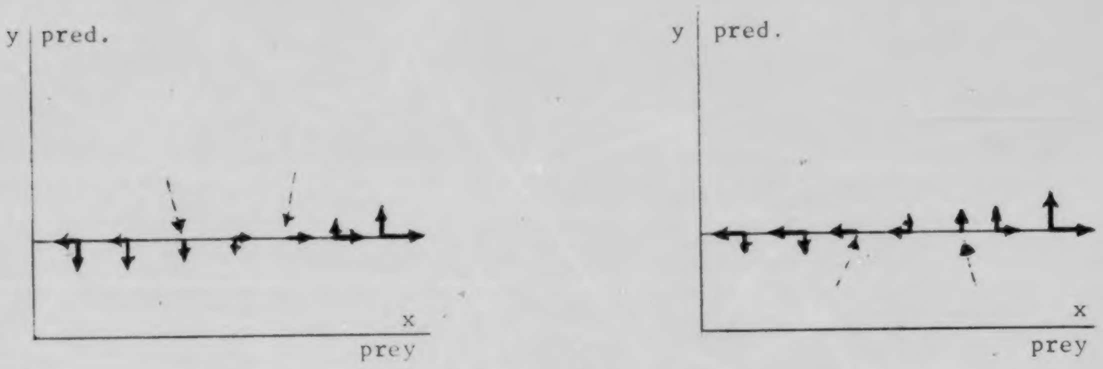
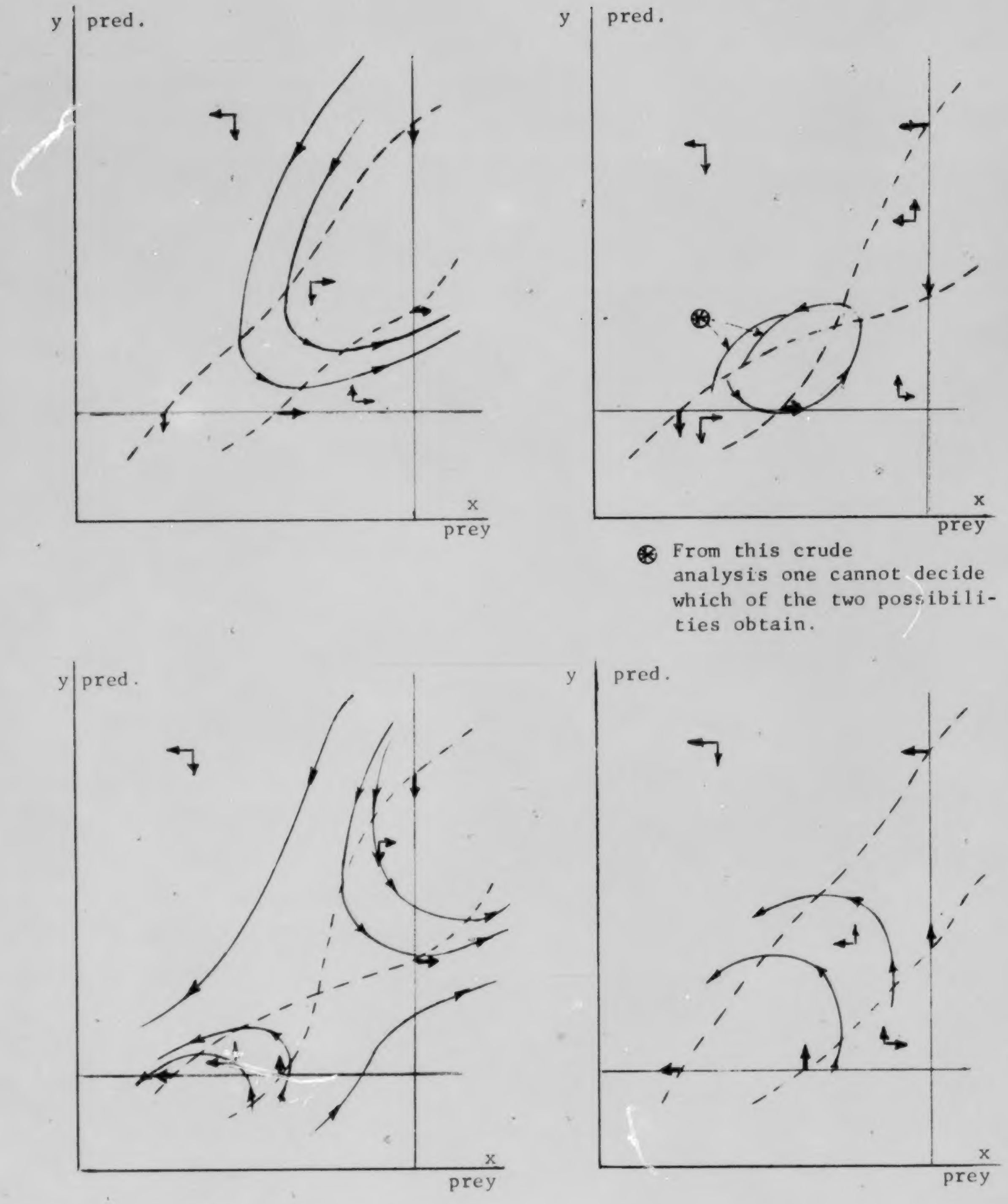


Figure 2.3,3

It should be said that in each of these diagrams, it is fairly unrealistic to consider portions of the phase plane close to the coordinate axes (x = 0 and/or y = 0): with numbers close to zero, there is a better chance for 'random influences' to step in, so the deterministic approach is not as valid here.

In conclusion, we find there are at least four possible cases of relative positions of the vertical and horizontal isoclines. These four possibilities are represented in the diagrams of Figure 2.3,4. The two lines that occurred in Figures 2.3,2 and 2.3,3 are repeated here, along with the crucial arrows located at the points indicated by the broken arrows in the two preceding figures: they indicate points of the 'horizontal' and 'vertical' isoclines.

Discussion Question 2.3,2. The reader should explain those arrows that are not on the isoclines.


y
pred.
x
prey
y
pred.
x
prey
y
pred.
x
prey
y
pred.
x
prey

⊛ From this crude analysis one cannot decide which of the two possibilities obtain.

Figure 2.3,4

Discussion Question 2.3,3. Two of the four possibilities depicted in Figure 2.3,4 are biologically unacceptable: which two? Why? The other two seem to suggest oscillations, damped or not, one (strong 'damping') or more. Show how the diagrams in Figure 2.3,4 lead you to make these statements.

Question 2.3,4. Comment on the apparent nature of the singular points, if any, in Figure 2.3,4.

Discussion Question 2.3,5. The reader should ponder the following point: the unspecific, qualitative arguments presented so far have not given any unambiguous clue as to the detailed location of certain parts of the isoclines in Figure 2.3,4. For instance, in the first and fourth diagrams the isoclines might yet intersect in the first quadrant: close to the x-axis; or they could even have two points of intersection in the portion depicted in Figure 2.3,4; as follows:

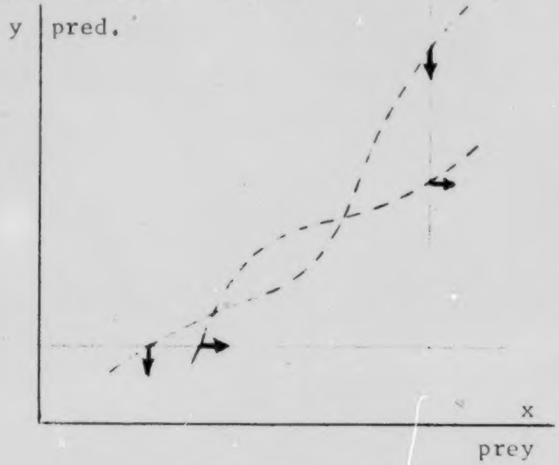


Figure 2.3,5

Inquire into the consequences for the phase portrait by the quick-and-dirty method. For any given real life ecosystem, of course, observations should be used to choose between such alternatives.

Remark 2.3,2. Rosenzweig (1969, Figure 1; reproduced in Figure 7-5 of Leigh, 1971) has used actual experimental data (from Huffaker, 1958) to construct a diagram very similar to those in our Figure 2,3.4. This is quite interesting for at least two reasons: (a) it shows that in this case a close parallel exists between mathematical and biological thinking, and (b) it illustrates a point that many mathematicians have trouble accepting, namely the fact that real life does not proceed in as clear-cut a manner as mathematicians would like to see: there are several arrows in Rosenzweig's diagram that are pointing the 'wrong' way, no doubt because many factors influence the real life situation that are not incorporated in the model.

3. MUTUALISM AND COMPETITION

3.1. Preliminary remarks.

Mutualism (e.g., see p. 245 of W. C. Allee et al., 1945; or pp. 25 and 26 of R. M. May, 1973a), commensalism (e.g., see p. 243 of Allee, l.c.; or pp. 25 and 26 of May, l.c.) and competition (e.g., see pp. 25, 26 of May, l.c.) may be explained in one scheme as follows:

a) commensalism is a relation between two species in which one species profits and the other species neither profits nor suffers;

b) mutualism is a relation between two species in which both species profit;

c) competition is a relation between two species in which both species suffer.

Certainly this scheme is a simplification: it leaves out the infinite variations on a basic theme which live ecosystems abundantly provide; especially competition is a subject about which much more can be said (for a starter see pp. 368, seqq., of Allee, l.c.; for more literature references see section 3.3 below). However, the above descriptions are enough for the purpose of constructing the simple Volterra-type models.

3.2. Mathematical models in the Volterra tradition.

The reader is invited to regard this section as one big discussion question and to analyze the following models by the methods discussed previously (i.e., phase plane-isocline-arrow analysis and linearization). As before, the parameters in these models are all understood to be positive, with one exception: the a_{ii} may be read either to be equal to zero (no "crowding" between the individuals of species i), or to be positive ("crowding" obtains). The reader should also explain these models in terms of the definitions given in section 3.1:

commensalism $$\begin{Bmatrix} \dot{x} = x\,(\epsilon_1 - a_{11}x) \\ \dot{y} = y\,(\epsilon_2 + a_{21}x - a_{22}y) \end{Bmatrix} \qquad (3.2,1)$$

mutualism $$\begin{Bmatrix} \dot{x} = x\,(\epsilon_1 - a_{11}x + a_{12}y) \\ y = y\,(\epsilon_2 + a_{21}x - a_{22}y) \end{Bmatrix} \qquad (3.2,2)$$

competition $$\begin{Bmatrix} \dot{x} = x\,(\epsilon_1 - a_{11}x - a_{12}y) \\ y = y\,(\epsilon_2 - a_{21}x - a_{22}y) \end{Bmatrix} \qquad (3.2,3)$$

The reader will find that with all $a_{ii} = 0$ the behavior of these models is rather unrealistic (populations explode). When analyzing these models with all $a_{ii} > 0$, the reader should find that the above model for commensalism necessarily shows a point $S = (x,y)$ with $x > 0$, $y > 0$, such that all solution curves tend to S as time increases. The model for mutualism shows either the same situation or a population explosion of both species, depending on whether $(a_{11}/a_{12}) > (a_{21}/a_{22})$ or $(a_{11}/a_{12}) \leq (a_{21}/a_{22})$. The model for competition shows four possibilities, depending on the disposition of the two ("horizontal" and "vertical" isoclines: all solutions tend (i) to a point in the interior of the first quadrant, or (ii) to a point on the x-axis, or (iii) to a point on the y-axis, or (iv) some solutions tend to a point on the x-axis, others to a point on the y-axis. The reader should find the inequalities that must be satisfied by the parameters to provide the conditions for each one of these four possibilities. The reader should make a sustained effort to make some biological sense out of the dichotomizing inequalities that govern the mutualism and competition models. Generally speaking they have to do with each species helping or hurting itself less or more than the other one.

The reader should also investigate whether he would want to improve on these models in some way reminiscent of the discussions in section 2.5 or in some other way.

Remark 3.2,1. Some people react to the results concerning competition with the feeling that they are unrealistic. There is no reason for this. In fact, there is ample experimental confirmation of the nature of these results. There are ecosystems, both in the laboratory and outside, that have been observed to exhibit the behavior predicted for some parameter values, viz., one species of the two disappears and the other stays. The behavior predicted by the above possibility (iv) has been extensively investigated in the Tribolium experiments (e.g., see Mertz, 1972, or Neyman, Park and Scott, 1956): the observations show that the fate of the Tribolium system depends on the initial ratio of the densities of the two species; roughly speaking, the species that is more numerous at the beginning stands a better chance of surviving. The above model goes a long way in explaining this behavior: just stochastize it, in the spirit of Discussion Question 1.6,6.

Question 3.2,1. The reader should investigate the behavior of this alternative model for competition:

$$\left\{\begin{aligned} \dot{x} &= x\,(\epsilon_1 - a_{11}\log_x - a_{12}\log y) \\ \dot{y} &= y\,(\epsilon_2 - a_{21}\log_x - a_{22}\log y) \end{aligned}\right\}. \qquad (3.2,4)$$

He will find that there is much qualitative similarity between the phase portrait of equation (3.2,3) and the phase portrait of equation (3.2,4). Here, too, there are four different possibilities. A difference that has assumed some importance in discussions about the competitive exclusion principle in the position, in possibility (i) (coexistence of the two species), of the coexistence point relative to the two singular points on the coordinate axes. An extensive discussion of equation (3.2,4) (too extensive for the present purpose) can be found in Gomatam, 1974b.

3.3. The competitive exclusion principle.

The four cases that we could distinguish in the competition model have given rise to a famous concept that has led to numerous experimental and some theoretical investigations and that still commands much attention in ecology. As the reader has noticed, three of the four possible cases of the Volterra model for competition between two species indicate the survival of one species and the extinction of the other. In one of these three cases it is not always the same species that survives: which one does survive depends on which was the most numerous species to begin with ('most numerous' in a slightly more technical sense: the question is on which side of a certain separatrix the initial point is located). Curiously enough, the fourth case was historically disregarded, and especially after the publication of Gause (1934), who also added some experimental evidence, the so-called competitive exclusion principle (sometimes called Gause's principle) took a firm hold in the ecological literature. It does not seem to be within the scope of the present discussion to pursue in detail the ups and downs of this principle ("two competing species cannot coexist," where one has to carefully define what is 'competing' and what is 'coexist'). However, the history of this 'principle' could easily serve as a paradigm of some of the ways in which a mathematical model, even a crude one, may be eminently useful, as well as some of the things that should not be expected of a mathematical model for biological problems, at least not yet.

If one looks over part of the literature: Gause (1934), Hardin (1960), Ayala (1970), Gilpin and Justice (1972), Stewart and Levins (1973), Gomatam (1974b), and the literature quoted by these authors, then it becomes rapidly clear that an enormous amount of experimental work was elicited by the enouncement of this "principle," which in turn was originated by the above-mentioned mathematical model (equation 3.2,3). (Even though some utterances of the principle were made at an earlier date, the work really started going after Gause (1934), whose work was explicitly provoked by Volterra's publications.) In the earlier years most of this work tended to confirm the 'principle,' but as ecologists probed deeper into the concept of competition, they started distinguishing situations in which the principle might not apply, either because of conceptual problems (such as: "Can this situation be labeled competition?") or because of features of the ecosystem's environment. Many questions came up: genetical differences, partitioning of the physical space into different habitats, larvae having needs different from adults, degree of competition (one resource shared, others not), etc. The eventual result seems to be that the original model (equation (3.2,3)) is seldom valid or applicable in a numerical sense (though valid in a broad, qualitative way), and that the concept of competition has branched out into a host of (sub-) concepts. But to have elicited all this work is not a bad performance for one little old model.

Remark 3.3,1. One of the points that drew attention recently was illustrated in Question 3.2,1: observations have been made of coexisting competing species, conforming to the positions of singular points as described by equation (3.2,4), rather than as described by equation (3.2,3); cf. Gilpin and Justice, 1972.

4. SOME REMARKS ON N-SPECIES SYSTEMS

Models for n-species in the Volterra tradition have the following appearance:

$$\dot{x}_i = x_i(\epsilon_i + \sum_{j=1}^{n} a_{ij}x_i),$$

where the ϵ_i may be positive or negative and where the a_{ij} may be positive or negative. Some special cases of this general set-up have been studied more than others.

Let species 1 prey on species 2, which in turn preys on species 3, which in turn ..., all the way down to species n who is prey to species n-1, without itself preying on any other species (note that this last point is sort of unrealistic; in a little while we'll receive heavier punishment for it than you might expect). In the tradition of equations (2.1,2) Volterra modeled such linear chains of predators by

$$\left.\begin{array}{l}
\dot{x}_1 = x_1(-\gamma_1 + \alpha_{12}x_2) \\
\dot{x}_2 = x_2(-\gamma_2 - \alpha_{21}x_1 + \alpha_{23}x_3) \\
\dot{x}_3 = x_3(-\gamma_3 - \alpha_{32}x_2 + \alpha_{34}x_4) \\
\dot{x}_4 = x_4(-\gamma_4 - \alpha_{43}x_3 + \alpha_{45}x_5) \\
\dot{x}_5 = x_5(+\gamma_5 - \alpha_{54}x_4)
\end{array}\right\}, \qquad (4,2)$$

where for convenience' sake we have taken $n = 5$ and where all γ_i and all α_{ij} are positive (see d'Ancona, 1954). By a change of scale one can always achieve that the matrix of the $\pm\alpha_{ij}$ is skew-symmetric; see Question 4,1 below. So we shall assume that this is already the case. Then the following argument is well-represented in the literature. Let $(p_1, p_2, p_3, p_4, p_5)$ be the point in $(x_1, x_2, x_3, x_4, x_5)$-space at which the 5 right-hand members of equation (4,2) have the value zero with all $p_i \geqq 0$. Let us write δ_i for γ_i $(i = 1, 2, 3, 4)$ and δ_5 for $-\gamma_5$. Let us write a_{ij} for α_{ij} $(i < j)$ and for $-\alpha_{ij}$ $(i > j)$. Let A be the matrix of the a_{ij}, A being skew-symmetric (see above); then the vector p satisfies the equation

$$A\,p = \delta \qquad (4,3)$$

The equations (4,2) can now be written

$$\dot{x}_i/x_i = \sum_j a_{ij}(x_j - p_j)$$

or, putting $x_i = p_i e^{v_i}$ (p_i constants, v_i functions of time t, as are x_i),

$$\dot{v}_i = \sum_j a_{ij}(e^{v_j} - 1)p_j \tag{4,4a}$$

From (4,4a) the reader will be able to derive the following two equations:

$$\sum_i \dot{v}_i p_i e^{v_i} = \sum_i \sum_j p_i e^{v_i} a_{ij} p_j (e^{v_j} - 1) \tag{4,4b}$$

$$\sum_i \dot{v}_i p_i = \sum_i \sum_j p_i a_{ij} p_j (e^{v_j} - 1) \tag{4,4c}$$

Now $\sum_i \sum_j q_i a_{ij} q_j = 0$ because the matrix A is skew-symmetric. Therefore the reader will be able to prove

$$\sum_i \dot{v}_i p_i (e^{v_i} - 1) = 0, \tag{4,4d}$$

that is

$$\frac{d}{dt} \sum_i p_i (e^{v_i} - v_i) = 0$$

$$\sum_i p_i (e^{v_i} - v_i) = \text{constant}, \tag{4,4e}$$

which the reader can easily rewrite in terms of the x_i. As a result, we have found a function of the x_i which remains constant through time: a so-called 'constant of motion,' or 'first integral.' This result is an n-species parallel to equation (2.1,6b).

The existence of this 'constant of motion' (whose biological interpretation seems to be out of reach) has enticed several people with a background in physics (e.g., Goel, Maitra and Montroll, 1971, where more references are cited) to apply the same methods to the Volterra model (4,2) as are used in the study of conservative systems in theoretical and statistical mechanics. There are many objections to this procedure. In the last analysis they all derive from one basic objection: the model (4,2) suffers from a serious lack of realism; for instance see the review of the book by Goel, et al. in Science of 2 June, 1972, p. 1007. Every criticism listed in section 2.2 applies here, and then some. We'll indicate a few in some of the following questions so that the reader may have the feeling that he helps discover some objections personally. The point is of course that if the model that constitutes the basis for further theoretical work is faulty, then what value can be attached to theoretical conclusions from it, no matter how fancy the means by which these conclusions are reached?

Question 4,1. Show that in equation (4,2) one can find a unique set of proportions $\beta_1 : \beta_2 : \beta_3 : \beta_4 : \beta_5$ such that for any set of α_{ij}

$$\alpha_{12} = A_{12}/\beta_1, \quad \alpha_{21} = A_{12}/\beta_2, \quad \alpha_{23} = A_{23}/\beta_2, \quad \alpha_{32} = A_{23}/\beta_3,$$

$$\alpha_{34} = A_{34}/\beta_3, \quad \alpha_{43} = A_{34}/\beta_4, \quad \alpha_{45} = A_{45}/\beta_4, \quad \alpha_{54} = A_{45}/\beta_5,$$

where the same A_{ij} occurs in connection with α_{ij} and α_{ji}. If subsequently the $\beta_i x_i(t)$ are put equal to $X_i(t)$, then a set of equations is obtained with a skew-symmetric matrix, as promised. It has been suggested that the $X_i(t)$ measure the species i $(i = 1, 2, 3, \ldots)$ in 'comparable' units. Do you like this suggestion?

Question 4,2. Show that as soon as species 1 would not only prey on species 2, but also on species 3, no set of proportions $\beta_1 : \beta_2 : \beta_3 : \beta_4 : \beta_5$ can be found anymore that would give us a system of equations with a skew-symmetric matrix, unless the coefficients α_{ij} satisfy some very precise conditions (in the form of strict equalities). It seems highly unlikely that many ecosystems exist that would satisfy such equalities. This means that as soon as the predator-prey relations no longer form a linear chain $(1 \to 2 \to 3 \to 4 \to 5)$, but have some extra link (e.g. $\begin{tikzcd} 1 \arrow[r] \arrow[rr, bend left] & 2 \arrow[r] & 3 \arrow[r] & 4 \arrow[r] & 5 \end{tikzcd}$), then a constant of motion would not exist, generally speaking.

Question 4,3. Show that as soon as among our species there would be two that compete against each other, or that are mutually advantageous to each other, we can't possibly have a skew-symmetric matrix, not even if the chain of relations is linear (e.g., $1 \to 2 \to 3 \to 4 — 5$, with 4 and 5 related by competition or mutualism).

Question 4,4. Show that as soon as logistic crowding terms are introduced in even one species, no amount of scale changing can achieve a skew-symmetric matrix. (Remember: a matrix A is skew-symmetric iff $A = -A^T$).

Question 4,5. So a linear chain of predator-prey relations is the only case for which it has been proven that a skew-symmetric matrix can be achieved.

Question 4,6. Our derivation (just like the ones in the literature) starts with equations (4,2) and (4,3), and vitally depends on the vector $\vec{p}$: all its components p_i must be strictly positive, or there is no way to make the substitution $x_i = p_i e^{v_i}$, which alone makes the transition to equation (4,4a) possible. The interesting thing is that the reader will be able to prove that for $n = 3, 5, 7, \ldots,$ in general for odd n, no such vector $\vec{p}$ will be forthcoming; again unless the a_{ij} and the δ_i satisfy certain very special relations: for odd n any skew-symmetric $n \times n$ matrix is singular. So the theory would seem to work only for ecosystems consisting of an even number of species interconnected by a linear predator-prey chain! This is the heavy punishment for lack of realism which we promised in the second paragraph of section 4.

Question 4,7. The biologically unacceptable fact that according to the model only even numbers of species could live in a steady state, evaporates if one adds plants and soil. This was proposed by Poletaev (1966) and E̊man (1966), whose work was quoted in section 2.2 (in the paragraph following Discussion Question 2.2,9). A very simple example of a model of such a biogeocoenotic system is provided by the following equations:

$$\left.\begin{array}{ll}\text{(predators)} & \dot{x} = x\,(-\gamma_1 + \alpha_{12}y) \\ \text{(prey)} & \dot{y} = y\,(-\gamma_2 - \alpha_{21}x + \alpha_{23}z) \\ \text{(plants)} & \dot{z} = z\,(-\gamma_3 - \alpha_{32}y + \alpha_{34}u) \\ \text{(soil nutrients)} & \dot{u} = \gamma_{41}x + \gamma_{42}y + \gamma_{43}z - \alpha_{43}u\,z \end{array}\right\} \quad (4,5)$$

Try and recover the thinking behind every term in equations (4,5). Investigate whether these equations admit an equilibrium point with x, y,

z, and u all strictly positive, and do the same for a model with one or more species in it. Would the parameters need to satisfy any special conditions in order that such an equilibrium point may exist? The equations (4,5) are of course crude (no 'crowding' terms, for instance). Specifically, cross-products such as $\alpha_{12}x\,y$ can be given a semblance of credibility by a chemical kinetics type of argument: animals move around and collide (as in collision theory); but a cross-product such as $\alpha_{43}u\,z$ is even harder to defend: plants do not move around. However, the idea of drawing plants and soil into the model is basically sound, obviously, and it is interesting that this is enough to get rid of the odd-even paradox.

Question 4.8. The reader can now construct a large number of Volterra (or even fancier) models, to represent one predator species living from two prey species (competing or mutualistic), or to represent two predator species competing for one prey, etc. Fishery models may be constructed with Man as the predator, etc. The possibilities are endless, but it is rather difficult to do analytical mathematical work on these models. One will often have to take recourse in computers. One many gain some insight in the severity of the mathematical problems by looking at Coppel (1966), who confesses that a satisfactory classification of systems of two differential equations with quadratic terms has not even been achieved yet. And these equations are of the oversimplified type that present research is leaving more and more behind! Therefore, further development of this field does very much depend on numerical methods, although there will always be room for clever qualitative analysis of these differential systems. To single out meaningful problems, mathematicains should consult extensively with interested ecologists.

APPENDIX A

A.1. Phase plane, isoclines, arrows.

Most mathematical models in this module are autonomous systems of ordinary differential equations:

$$\left\{\begin{aligned} \dot{x}(t) &= \varphi(x(t),\ y(t)) \\ \dot{y}(t) &= \psi(x(t),\ y(t)) \end{aligned}\right\} \tag{A.1,1}$$

that is, the values of the derivatives $\dot{x}(t)$ and $\dot{y}(t)$ at any time t are uniquely determined by the values of the 'state variables,' $x(t)$ and $y(t)$, regardless of the time t at which these values are attained.

> Question A.1,1. Is the equation $\dot{x}(t) = x(t)$ autonomous? Is the equation $\dot{x}(t) = t\,x(t)$ autonomous?

The discussion in the present appendix will concentrate on autonomous systems.

In several ways the study of the properties of solutions of autonomous systems is simpler than of non-autonomous systems, especially in the case of dimension 2: two variables, say x and y, both to be determined as functions of t. For openers, we'll take two autonomous systems which can be solved explicitly, so that we may compare the results of alternative approaches with the correct answers. Consider

$$\text{I}\quad \begin{aligned} \dot{x} &= a, \quad a > 0 \\ \dot{y} &= b, \quad b > 0 \end{aligned} \qquad \text{and} \qquad \text{II}\quad \begin{aligned} \dot{x} &= a\,x, \quad a > 0 \\ \dot{y} &= b\,y, \quad b > 0 \end{aligned} .$$

Their solutions, for $x(t_o) = x_o$, $y(t_o) = y_o$, are

$$\text{I}'\left\{\begin{aligned} x(t) &= x_o + (t-t_o)a \\ y(t) &= y_o + (t-t_o)b \end{aligned}\right\} \qquad \text{and} \qquad \text{II}'\left\{\begin{aligned} x(t) &= x_o\cdot e^{(t-t_o)a} \\ y(t) &= y_o\cdot e^{(t-t_o)b} \end{aligned}\right\},$$

as is easily checked by substituting I' and II' into I and II, respectively.

Each pair of equations such as (A.1,1) determines a family of curves in (t,x,y)-space. It is excellent training for the students to visualize how these curves are going in (t,x,y)-space. This is easy for system I, and a bit harder for system II:

Solutions I'. Choose numerical values for a, b, t_o, x_o and y_o and 'plot' the corresponding curve in 3-dimensional space. That is, for each value t_i of t imagine a plane perpendicular to the t-axis at the point $t = t_i$; in such a plane then plot the point with coordinates

$$x(t_i) = x_o + (t_i - t_o)a, \quad y(t_i) = y_o + (t_i - t_o)b.$$

All these points $(i = 1, 2, 3, \ldots)$ string along and (in the case of system I !) determine a straight line, tilted in space. Then choose different values for x_o and y_o, say x'_o and y'_o, and find another straight line, parallel to the first one. Or choose a different value for t_o, say t'_o, and use the same old x_o and y_o: you'll find yet another straight line, parallel to the first one. While you're plotting the points of all those lines in (t,x,y)-space, you may also plot their projections onto the (x,y)-plane (which is easy: the projection of (t_i,x_i,y_i) is (x_i,y_i)). The projection of the solution curve through the point (t_o,x_o,y_o) consists of all points (x_i,y_i) for which

$$x_i = x_o + (t_i - t_o)a, \quad y_i = y_o + (t_i - t_o)b.$$

These points all belong to the line (in the (x,y)-plane):

$$\frac{x - x_o}{a} = \frac{y - y_o}{b}, \quad \text{or} \quad (x - x_o)b = (y - y_o)a.$$

The slope of any such line at the point $(x,y) = (\xi,\eta)$ is obviously b/a, which ratio equals the value of $\dot{y}/\dot{x}$ as determined by equation I (at the point (ξ,η). Finally note that all solution curves in (t,x,y)-space that pass through any of the points (t_o,x_o,y_o) with fixed (x_o,y_o) and variable t_o (each value of t_o giving rise to one solution curve!) are parallel and have one and the same projection onto the (x,y)-plane. This final property is very much a consequence of system I being autonomous. It enables us to find out the states (x,y) through which the system is going to move or has moved (we'll call both together the 'life history of the system'), simply by looking at the projection of the pertinent solution curve(s): if the point (x_o,y_o) belongs to the system's life history, then the pertinent solution curve(s) should contain some point (t,x_o,y_o), and the pertinent projection should contain the point (x_o,y_o). The cylinder erected on any such projection with generators parallel to the t-axis is filled with congruent solution curves, which can be obtained from each other by suitable translations parallel to the t-axis.

Remark A.1.1. In the case of system I the 2-vector valued functions determined by the equations

$$(I') \begin{Bmatrix} x = x_o + (t - t_o)a \\ y = y_o + (t - t_o)b \end{Bmatrix}$$

are called solutions; the curves in (t,x,y)-space determined by the same equations are called solution curves (straight lines in this case); and the projections of these curves onto the (x,y)-plane are called orbits, trajectories, or paths; in the case of system I the orbits are determined by the equation

$$(x - x_o)b = (y - y_o)a,$$

which describes the life history of such systems; note that this equation is obtained from (I') by the elimination of t. Further, the (x,y)-plane itself is denoted as phase plane. Since the ratio $\dot{y}/\dot{x} = b/a$ has, for system I, the same value at every point of the phase plane, the concept of isocline (see below) degenerates here.

Solutions II'. Again choose numerical values for a, b, t_o, x_o, y_o and 'plot' the solution curves in (t,x,y)-space. The result is somewhat harder to visualize than it was for I', but it is quite feasible. And again, if we compare all solution curves that pass through points (t_o,x_o,y_o) with fixed (x_o,y_o) and variable t_o (each value of t_o giving rise to one solution curve!), we see that these curves have parallel directions at the points (t,x_o,y_o); in fact, these curves can be obtained from each other by suitable translations parallel to the t-axis. Thus we can use their common projection (viz., the orbit through (x_o,y_o)) to find out the life history of each solution curve that passes through some point (t,x_o,y_o) (t arbitrary, (x_o,y_o) fixed). As before, we find the equation of this orbit by eliminating t from the equations (II'):

$$\left(\frac{x}{x_o}\right)^b = \left(\frac{y}{y_o}\right)^a.$$

The reader should explore the shape of this curve, for the time being only for $0 < a < b$ (other cases later); simply solve for y and plot in the (x,y)-plane (note that all points of the curve are in the same quadrant as the point (x_o,y_o): why? This also follows from the equations II'). Also find the derivative of y as a function of x, at the point (ξ,η):

$$\frac{dy}{dx} = \frac{b}{a}\frac{\eta}{\xi}$$

(remember, the point (ξ,η) is on the orbit defined by $(x/x_o)^b = (y/y_o)^a$!) This equation for dy/dx can easily be obtained from the original system (II):

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{b}{a}\frac{\eta}{\xi} .$$

Remark A.1.2. Without first finding explicit expressions for the solutions I' of the system I or the solutions II' of the system II one can still obtain a fairly good insight in the shape of the orbits by plotting the slopes of the (tangent lines to the) orbits at a sufficiently large number of points in the phase plane.* This process can be made a lot easier by making use of isoclines. An isocline is the locus of all those points (x,y) in the phase plane at which the slope of the orbit passing through (x,y) has a given value.

Question A.1.2. Consider the system

$$\begin{Bmatrix} \dot{x} = x \\ \dot{y} = 2y \end{Bmatrix}$$

The slope of the orbit through the point (ξ,η) has, at that same point, the value

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{2\eta}{\xi} .$$

The locus of all points where that slope is +1 is given by the equation

$$\frac{2\eta}{\xi} = 1.$$

The locus of all points where that slope is +2 is given by

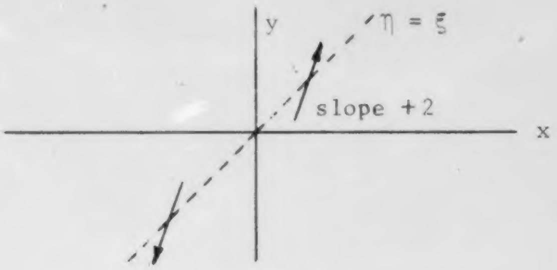
$$\frac{2\eta}{\xi} = 2.$$

So $\eta = \frac{1}{2}\xi$ is the isocline corresponding to the slope +1, etc. Find the isoclines corresponding to the slopes $\frac{1}{4}$, $\frac{1}{2}$, 3, $-\frac{1}{4}$, $-\frac{1}{2}$, -3. Then draw these isoclines and through some of their points draw (carefully!) short line segments which represent the tangent lines to the orbits through these points at these points. (These drawings should be on a fairly large scale in order to achieve reasonable accuracy). Now you can

---

* This is one of the alternative approaches mentioned in the beginning of section A.1.

sketch a few orbits since you have now enough information on their directions at various and sundry points. Note that the above-mentioned short line segments can be provided with arrows to indicate the direction in which time increases along the orbit (e.g., if (ξ,η) is in the first quadrant, then $\dot{x} = \xi > 0$ and $\dot{y} = 2\eta > 0$ and so at (ξ,η) an orbit moves up and to the right as time progresses; if (ξ,η) is in the third quadrant, it is the other way around; see diagram)



Also note that none of the above isoclines coincides with or even is tangent to an orbit!

Definitions. Before going on, we'll write three definitions.

Solutions. Given a system of (ordinary) (autonomous) differential equations

$$\dot{\vec{x}} = \vec{f}(\vec{x}) \quad \text{or} \quad \left\{ \begin{array}{l} \dot{x}_1 = f_1(x_1,x_2) \\ \dot{x}_2 = f_2(x_1,x_2) \end{array} \right\},$$

where x_1 and x_2 represent two unknown functions of t, a solution is a fully specified 2-vector valued function of t,

$$\vec{x} = \vec{\phi}(t) \quad \text{or} \quad \left\{ \begin{array}{l} x_1 = \phi_1(t) \\ x_2 = \phi_2(t) \end{array} \right\},$$

such that substitution of $\vec{\phi}(t)$ for $\vec{x}$ reduces the above differential equation to an identity in t. A curve in (t,x_1,x_2)-space that is defined by a solution $\vec{x} = \vec{\phi}(t)$ is called a solution curve.

Question A.1.3. Check the claim that II' is the solution of II, against the just given definition of solution.

Remark A.1,3. We have seen earlier that solutions depend on so-called initial conditions: if the value of the solution must be $\vec{x}_o$ at time t_o, then it is often useful to incorporate this condition into the notation:

$$\vec{x} = \vec{\Phi}(t;\ t_o,\vec{x}_o) \quad \text{or} \quad \left\{\begin{aligned} x_1 &= \Phi_1(t;\ t_o,x_{o1},x_{o2}) \\ x_2 &= \Phi_2(t;\ t_o,x_{o1},x_{o2}) \end{aligned}\right\}$$

where $\vec{\Phi}(t_o;\ t_o,\vec{x}_o) = \vec{x}_o$.

Orbits, trajectories, or paths are curves in the phase plane (here, the (x_1,x_2)-plane), which geometrically are projections of solution curves, and which analytically are determined either by elimination of t from the two above equations $x_1 = \Phi_1(t;\ \ldots)$ and $x_2 = \Phi_2(t;\ \ldots)$, or by regarding these two equations as the parametrization of some curve (viz., the orbit) in the (x_1,x_2)-plane.

Remark A.1,4. If the solutions of a differential system are available in explicit form, there is no special problem in studying the orbits. If the solutions are not available, then it may happen that the equation of the orbits can be found (see Question 2.1,3), although this is relatively rare. If this is not the case either, then we can choose between several alternative methods (and often use them all) for studying the orbits. The easiest of these is the quick-and-dirty isocline and arrow method. Its basis is the formula which we have already applied:

$$\frac{dx_2}{dx_1} = \frac{\dot{x}_2}{\dot{x}_1} = \frac{f_2(x_1,x_2)}{f_1(x_1,x_2)} ,$$

which gives the direction of the tangent line to any orbit at any point.

Isoclines. Given the differential system

$$\left\{\begin{aligned} \dot{x}_1 &= f_1(x_1,x_2) \\ \dot{x}_2 &= f_2(x_1,x_2) \end{aligned}\right\}$$

the curve which in (x_1,x_2)-plane is defined by

$$(\cos\alpha)\cdot f_2(\xi_1,\xi_2) = (\sin\alpha)\cdot f_1(\xi_1,\xi_2),$$

is called the isocline corresponding to the slope $\tan\alpha$; it is the locus of all points (ξ_1,ξ_2) such that the orbit through (ξ_1,ξ_2) has a tangent line at that point with slope $\tan\alpha$.

Remark A.1.5. In general an isocline is not an orbit; in general an orbit and an isocline have no common tangent lines at their points of intersection (though there are exceptions to this rule!). Among the continuum of isoclines two are of outstanding importance in most problems: the isocline corresponding to $\alpha = 0$ and the one corresponding to $\alpha = \pi/2$; that is, the locus of points where the orbits have a 'horizontal' direction (x_2 stationary; $\dot{x}_2 = 0$), and the locus of points where the orbits have a 'vertical' direction (x_1 stationary; $\dot{x}_1 = 0$). For examples and applications see section A.2 as well as many questions in the main part of the text of this module.

A.2. The quick-and-dirty isocline and arrow method.

At the beginning of this method stands the drawing of two isoclines, one corresponding to $\alpha = 0$ and the other to $\alpha = \pi/2$.

Question A.2.1. For the following systems of differential equations find the equations for both the $(\alpha = 0)$-isocline and the $(\alpha = \pi/2)$-isocline, and sketch them.

$$\text{I} \qquad \left\{\begin{array}{l} \dot{x} = +px + \beta y \\ \dot{y} = -\beta x + py \end{array}\right\}, \quad p > 0, \quad \beta > 0$$

$$\text{II} \qquad \left\{\begin{array}{l} \dot{x} = -px - \beta y \\ \dot{y} = +\beta x - py \end{array}\right\}, \quad p > 0, \quad \beta > 0$$

$$\text{III} \qquad \left\{\begin{array}{l} \dot{x} = \beta y \\ \dot{y} = -\beta x \end{array}\right\}, \quad \beta > 0$$

$$\text{IV} \qquad \left\{\begin{array}{l} \dot{x} = -\beta y \\ \dot{y} = +\beta x \end{array}\right\}, \quad \beta > 0$$

$$\text{V} \qquad \left\{\begin{array}{l} \dot{x} = -\lambda_1 x + y^2 \\ \dot{y} = -\lambda_2 y + x^2 \end{array}\right\}, \quad \lambda_1 > 0, \quad \lambda_2 > 0$$

$$\text{VI} \qquad \left\{\begin{array}{l} \dot{x} = x(+\gamma_1 - \delta x - py) \\ \dot{y} = y(-\gamma_2 + qx - \zeta y) \end{array}\right\} \quad \left\{\begin{array}{lll} \gamma_1 > 0, & \gamma_2 > 0, & \delta > 0 \\ \zeta > 0, & p > 0, & q > 0 \end{array}\right\}$$

In order to gain profitable information from these isoclines concerning the shape of the orbits we supplement our diagrams with a few easily derived items:

α) For each of the six above differential equation systems note that the value of $\dot{x}$ is negative on one side of the $(\alpha = \pi/2)$-isocline $(\alpha = \pi/2$ iff $\dot{x} = 0)$ and positive on the other side. Therefore one can draw a horizontal arrow pointing left wherever $\dot{x} < 0$ and a horizontal arrow pointing right wherever $\dot{x} > 0$. In fact, one can organize this in a systematic fashion. First, note that in each of the above six cases the two isoclines partition the phase plane into a number of subsets S: four in the first four cases, five in the fifth case and eleven in the sixth case. Second, there are always a few among those subsets S for which it is very easy to determine the sign of $\dot{x}$; e.g., for system I $\dot{x} > 0$ at any point in the first quadrant. Third, it is obvious (as long as the functions $\varphi(\ )$ and $\psi(\ )$ in

$$\left\{ \begin{array}{l} \dot{x} = \varphi(x,y) \\ \dot{y} = \psi(x,y) \end{array} \right\}$$

are continuous; and they are in all cases discussed in this module) that the sign of $\dot{x}$ must be the same at all points inside any given subset S: indeed, $\varphi(x,y)$, being continuous, cannot change sign without passing the value 0 and so the sign of $\dot{x}$ stays the same from one point to another point unless you have to cross the $(\dot{x} = 0)$-isocline once (or thrice, etc.) in order to get from the first point to the second. Fourth, it is helpful to, whenever you cross the $(\dot{y} = 0)$-isocline, put an arrow on that isocline, pointing left or right, as the case may be (note that the sign of $\dot{x}$ does not change when you cross the $(\dot{y} = 0)$-isocline). Fifth, again because of the continuity of $\varphi(\ )$, it is clear that the value of $\dot{x} = \varphi(x,y)$ approaches 0 as the point (x,y) approaches an $(\dot{x} = 0)$-isocline (one could express this by making the arrows shorter, but one may as well just keep it in mind for future reference).

Question A.2.2. In each of the six diagrams that resulted from the preceding question draw all the horizontal arrows, → or ←, according to the above specifications.

Question A.2.3. In the following system $\dot{x}$ never changes sign, not even when crossing the $(\dot{x} = 0)$-isocline. Why not?

$$\dot{x} = (px + \beta y)^2$$

$$\dot{y} = -\beta x + py$$

β) Go over every sentence in the above paragraph marked (α) and see how it is changed when our concern is the sign of $\dot{y}$ instead of $\dot{x}$.

Question A.2,4. In each of the six diagrams that you worked on in Question A.2,2, draw all the vertical arrows, ↑ or ↓ , according to the specifications listed under (β).

γ) Finally, when in a certain subset S you find the arrows ↑ and →, this means that in that subset the orbits move up and to the right

as time passes on. Similarly for the other three possibilities. Using this information, including the directions in which the isoclines must be crossed, one can start sketching the orbits: flattening them as you approach a ($\dot{y} = 0$)-isocline, making them steeper as you approach an ($\dot{x}=0$)-isocline. (Why?).

Question A.2,5. Do all the sketching of orbits you can in the six diagrams that you have been working on in Question A.2,4. Note that certain conclusions suggest themselves strongly from these sketches, although one would try and find confirmation from some more precise argument. For example,

- for system I, the orbits seem to spiral outward from the origin;
- for system II, the orbits seem to spiral inward toward the origin;
- for systems III and IV, the orbits seem to go around in circles about the origin (draw some isoclines with different α);
- for system V, the origin seems to be an attractor (no spirals!), and the point $\left((\lambda_1\lambda_2^2)^{1/3}, (\lambda_1^2\lambda_2)^{1/3}\right)$ seems to be a saddlepoint;
- for system VI, two cases are to be distinguished: the slanted parts of the isoclines intersecting in the first, or in the fourth, quadrant; in the first case the origin and the point $(\gamma_1/\delta, 0)$ seem to be saddle-points, the point $(o, -\gamma_2/\zeta)$ a repulsor, and the point in the interior of the first quadrant an attractor (with or without spirals; that is not too clear).

A.3. *Linear systems*.

The quick-and-dirty isocline and arrow method is *global*. One can usually sketch orbits both far from and close to the origin or any singular point.

Remark A.3,1. A *singular point* is any point in the phase plane that is an orbit in itself: if a system would be at any singular point, it would stay there (mathematically speaking; however, if the singular point is 'unstable,' then an analog computer would eventually get away from it, either because it would not ever get precisely there to begin with, or because of electronic noise). Singular points are found by requiring that both $\dot{x}$ and $\dot{y}$ equal zero, $\varphi(x,y) = 0 = \psi(x,y)$, cf. equation (A.1,1).

The easiest way to confirm or refute the suggestions made by the quick-and-dirty method, and also to supplement those suggestions, is *linearization*. the basic idea of linearization is to study the *local* behavior of the orbits of a given system in the neighborhood of any of its singular points by studying the uniquely defined *linear system* which under certain conditions *approximates* the *given system*.

Remark A.3,2. Linearization is not always possible, e.g. at the origin in the system

$$\left\{ \begin{aligned} \dot{x} &= a_{11}x^2 + 2\,a_{12}xy + a_{22}y^2 \\ \dot{y} &= b_{11}x^2 + 2\,b_{12}xy + b_{22}y^2 \end{aligned} \right\}.$$

Linearization is not always decisive (see below), and linearization gives only local results. Thus limit cycles, for instance, need other methods to prove or disprove their existence, but such will be outside the scope of this module.

In order to understand this linearization method, we had better first take a look at *linear systems* (their local behavior implies their global behavior). We will restrict ourselves to *two-dimensional*, autonomous systems:

$$\dot{\vec{x}} = A\vec{x} \qquad \text{(A.3,1)}$$

or

$$\left\{ \begin{aligned} \dot{x} &= a_{11}x + a_{12}y \\ \dot{y} &= a_{21}x + a_{22}y \end{aligned} \right\} \qquad \text{(A.3,1a)}$$

Obviously (why?) this system has only one singular point, the origin (that is, provided the matrix A is nonsingular). A linear system with singular point $\vec{p} = (p,q)$ is

$$\dot{\vec{x}} = A(\vec{x} - \vec{p})$$

or

$$\left\{ \begin{aligned} \dot{x} &= a_{11}(x-p) + a_{12}(y-q) \\ \dot{y} &= a_{21}(x-p) + a_{22}(y-q) \end{aligned} \right\} \qquad \text{(A.3,1b)}$$

The behavior of the second system follows directly from that of the first. So we will worry only about (A.3,1). Note that we are restricting ourselves to autonomous systems, so A is a constant matrix. Also note that in general A is not symmetric.

The basic strategy for studying the solutions and orbits of equation (A.3,1) consists in studying only a few special cases (i.e., a few special matrices A) and showing that the general case can be reduced to one of the special cases by means of a suitable linear transformation--much like the study of conics and quadrics in analytic geometry. The linear transformations work as follows. Let P be a constant, nonsingular matrix. Define the unknown vector-valued function $\vec{z}(t)$ in terms of the unknown vector-valued function $\vec{x}(t)$ by

$$\vec{z}(t) = P\,\vec{x}(t) \;\Leftrightarrow\; \vec{x}(t) = P^{-1}\,\vec{z}(t) \;\Leftrightarrow\; \dot{\vec{z}}(t) = P\,A\,P^{-1}\,\vec{z}(t);$$

and all we need to do is to choose P so that $P\,A\,P^{-1}$ is one of a few special matrices. The choice of the matrix P and the resulting type of special matrix $P\,A\,P^{-1}$ depend on the eigenvalues of A.

Definition. The eigenvalues (or characteristic values, or latent roots) of any matrix A are defined as the roots of the polynomial equation

$$\det(A - \lambda I) = 0, \quad \text{or}$$

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0, \quad \text{or} \quad \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0.$$

In all applications in this module the elements of the matrix A are real. Therefore the eigenvalues are either real or conjugate complex. In fact, we will distinguish between the following cases:

1°) Eigenvalues real and different	both eigenvalues nonzero.
2°) Eigenvalues real and equal	
3°) Eigenvalues pure imaginary and different	
4°) Eigenvalues complex and different	
5°) One eigenvalue zero, one nonzero	one eigenvalue zero.
6°) Both eigenvalues zero	both eigenvalues zero.

Representatives of these six possibilities will now be discussed, roughly in order of complexity.

6°) Here we have to distinguish between two cases:

6°)A

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

Every point of the phase plane is a singular point!

6°)B

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

This is a more interesting case. The reader will have no difficulty in showing that every point in the z_1-axis ($z_2 = 0$) is a singular point; that all orbits outside the z_1-axis are parallel to the z_1-axis, moving to the right (as $t \uparrow +\infty$) above the z_1-axis, and to the left below. The solution functions are obviously defined by

$$\vec{z}(t) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \vec{z}(0)$$

Prove!

5°)

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

Every point of the z_2-axis in the phase plane is a singular point; the orbits are horizontal half-lines (do apply also the isocline and arrow method!).

3°) $$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ -\beta & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \text{or} \quad \left\{ \begin{aligned} \dot{z}_1 &= \beta z_2 \\ \dot{z}_2 &= -\beta z_1 \end{aligned} \right\}$$

The orbits can be obtained very quickly by multiplying the first equation by z_1 and the second by z_2. Then add

$$z_1 \dot{z}_1 + z_2 \dot{z}_2 = 0$$

So the orbits are circles around the origin:

$$z_1^2(t) + z_2^2(t) = z_1^2(t_o) + z_2^2(t_o).$$

Alternatively, we can study the value of $dz_2/dz_1 = -z_1/z_2$; so the slope of the orbit at the point (z_1,z_2) is orthogonal to the slope of the line through the origin and the point (z_1,z_2) (brush up that analytic geometry!).

Alternatively we can combine the two equations into one:

$$\ddot{z}_1 = -\beta^2 z_1,$$

with the general solution,

$$z_1(t) = \gamma_1 \cos \beta(t-t_o) + \gamma_2 \sin \beta(t-t_o),$$

from which z_2 follows as $\beta^{-1} \dot{z}_1(t)$:

$$z_2(t) = \gamma_2 \cos \beta(t-t_o) - \gamma_1 \sin \beta(t-t_o)$$

Introducing initial conditions

$$z_1(t_o) = z_{10}, \quad z_2(t_o) = z_{20},$$

we find

$$\left\{ \begin{aligned} z_1(t) &= z_{10} \cos \beta(t-t_o) + z_{20} \sin \beta(t-t_o) \\ z_2(t) &= z_{20} \cos \beta(t-t_o) - z_{10} \sin \beta(t-t_o) \end{aligned} \right\}$$

or

$$\vec{z}(t) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \vec{z}(t_o)$$

with obvious notation. The reader should check these results against the isocline and arrow method and convince himself that they are in perfect harmony.

4°) $$\begin{bmatrix} \dot z_1 \\ \dot z_2 \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \text{or} \quad \left\{ \begin{matrix} \dot z_1 = \alpha\, z_1 + \beta\, z_2 \\ \dot z_2 = \beta\, z_1 + \alpha\, z_2 \end{matrix} \right\}$$

Here we need to make a clever change of variables:

$$\vec z(t) = e^{\alpha t}\, \vec w(t), \quad \vec z(t_o) = e^{\alpha t_o}\, \vec w(t_o).$$

This reduces our equation to the form we already know:

$$\left\{ \begin{matrix} \dot w_1 = & \beta\, w_2 \\ \dot w_2 = -\beta\, w_1 & \end{matrix} \right\}.$$

Making the necessary substitution the reader will now easily find that

$$z(t) = e^{(t-t_o)\alpha} \begin{bmatrix} \cos\beta(t-t_o) & \sin\beta(t-t_o) \\ -\sin\beta(t-t_o) & \cos\beta(t-t_o) \end{bmatrix} \vec z(t_o)$$

Defining $r(t)$ as the distance between the origin and the running point $\vec z(t)$, the reader will also find $\langle r^2(t) = \vec z^T(t)\, \vec z(t) \rangle$:

$$r(t) = e^{(t-t_o)\alpha}\, r(t_o),$$

which shows that the point $\vec z(t)$ in the phase plane describes a spiral, inward as $t \uparrow \infty$ when α is negative, outward as $t \uparrow \infty$ when α is positive.

Question A.3,1. An alternative way of handling (4°) is to change to polar coordinates

$$\left\{ \begin{matrix} z_1(t) = r(t)\cos\varphi(t) \\ z_2(t) = r(t)\sin\varphi(t) \end{matrix} \right\} \quad \text{or} \quad \left\{ \begin{matrix} z_1(t_o) = r(t_o)\cos\varphi(t_o) \\ z_2(t_o) = r(t_o)\sin\varphi(t_o) \end{matrix} \right\}$$

This approach leads to

$$\left\{ \begin{matrix} \varphi(t) = \varphi(t_o) - (t-t_o)\beta \\ r(t) = e^{(t-t_o)\alpha}\, r(t_o) \end{matrix} \right\}.$$

Question A.3,2. Let the reader apply the isocline and arrow method (both for $\alpha > 0$ and for $\alpha < 0$; also for $\beta > 0$ and for $\beta < 0$) and demonstrate that its results are in perfect harmony with the above.

2°)A.
$$\begin{bmatrix} \dot z_1 \\ \dot z_2 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \text{or} \quad \left\{ \begin{matrix} \dot z_1 = \lambda z_1 \\ \dot z_2 = \lambda z_2 \end{matrix} \right\}$$

The solutions follow immediately:

$$\left\{ \begin{matrix} z_1(t) = e^{(t-t_o)\lambda} z_1(t_o) \\ z_2(t) = e^{(t-t_o)\lambda} z_2(t_o) \end{matrix} \right\} .$$

Question A.3,3. Show that the point $\vec{z}(t)$ is in the same quadrant as $\vec{z}(t_o)$ for all t between $-\infty$ and $+\infty$.

Question A.3,4. Show that the orbits are straight half-lines, and that the direction of movement (away from or toward the origin as $t \uparrow \infty$) depends on the sign of λ.

Question A.3,5. Find the slope of the orbit at the point (z_1, z_2).

Question A.3,6. Show that in this case orbits and isoclines coincide.

2°)B.
$$\begin{bmatrix} \dot z_1 \\ \dot z_2 \end{bmatrix} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \text{or} \quad \left\{ \begin{matrix} \dot z_1 = \lambda z_1 + z_2 \\ \dot z_2 = \lambda z_2 \end{matrix} \right\}$$

Question A.3,7. The isocline and arrow method is quite suggestive for this system. The horizontal axis is both orbit and isocline, but no other orbit coincides with any other isocline. Do sketch isoclines and orbits for this system. Distinguish between λ positive and λ negative.

The second equation of this system is easy to solve: $z_2(t) = e^{(t-t_o)\lambda} z_2(t_o)$. Plug this solution into the first equation, multiply the result by $e^{-\lambda t}$, and show that the derivative of $e^{-\lambda t} z_1(t)$ shall be equal to the derivative of $e^{-\lambda t_o} z_2(t_o) \cdot t$, from which result it easily follows that

$$\left\{\begin{aligned} z_1(t) &= e^{(t-t_o)\lambda} \langle z_{10} + z_{20}\cdot(t-t_o)\rangle \\ z_2(t) &= e^{(t-t_o)\lambda} z_{20} \end{aligned}\right\}$$

Question A.3,8. Analyse the orbits defined by this equation. First take negative λ, and show that for orbits with $(z_{10},z_{20}) = \langle z_1(t_o), z_2(t_o)\rangle$ in the upper half-plane

$z_1(t)$ reaches a maximum

at the point where the orbit intersects the line $\lambda z_1 + z_2 = 0$, which is the $(\dot{z}_1 = 0)$-isocline (that makes sense!). How does this change either when (z_{10},z_{20}) is in the lower half-plane or when λ is positive?

1°)
$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \text{or} \quad \left\{\begin{aligned} \dot{z}_1 &= \lambda_1 z_1 \\ \dot{z}_2 &= \lambda_2 z_2 \end{aligned}\right\}$$

Here we distinguish between the case where λ_1 and λ_2 have the same sign, and the case where the two have different signs.

1°)A. $\lambda_1 < 0$ and $\lambda_2 < 0$. The solutions are

$$\left\{\begin{aligned} z_1(t) &= e^{(t-t_o)\lambda_1} z_1(t_o) \\ z_2(t) &= e^{(t-t_o)\lambda_2} z_2(t_o) \end{aligned}\right\}$$

Again, the point $\vec{z}(t)$ is for all t in the same quadrant as $\vec{z}(t_o)$. The point $\vec{z}(t)$ will approach the origin as $t \uparrow \infty$. For all but two orbits (besides the origin itself) the asymptotic direction (as $t \uparrow \infty$) will be

$$\left\{\begin{aligned} &\text{the } z_1\text{-axis iff } -|\lambda_2| < -|\lambda_1| < 0 \\ &\text{the } z_2\text{-axis iff } -|\lambda_1| < -|\lambda_2| < 0 \end{aligned}\right\}.$$

Question A.3,9. Find the equation of the orbits by eliminating t. Then express z_2 in terms of z_1 and show that the slope at the origin is given by

$$\begin{cases} \dfrac{dz_2}{dz_1} = 0 \quad \text{if} \quad 0 < |\lambda_1| < |\lambda_2| \\ \\ \dfrac{dz_1}{dz_2} = 0 \quad \text{if} \quad 0 < |\lambda_2| < |\lambda_1| \end{cases}$$

Question A.3,10. Identify the two orbits that are different (hint: take $z_{10} = 0$, or $z_{20} = 0$).

1°)B. $\lambda_1 > 0$ and $\lambda_2 > 0$. The solutions are given by the same formulae, but now the point $\vec{z}(t)$ will go away from the origin as $t \uparrow \infty$, and approach the origin as $t \downarrow -\infty$. For all but 2 orbits (besides the origin itself) the asymptotic direction (as $t \downarrow -\infty$) will be

$$\begin{Bmatrix} \text{the } z_1\text{-axis iff } 0 < \lambda_1 < \lambda_2 \\ \text{the } z_2\text{-axis iff } 0 < \lambda_2 < \lambda_1 \end{Bmatrix}.$$

Question A.3,11. Prove this and find the two exceptional orbits.

1°)C. $\lambda_1 < 0 < \lambda_2$. The solutions are given by the same formulae as in 1°)A, but the result of the elimination of t is now

$$z_2^{|\lambda_1|} \; z_1^{|\lambda_2|} = z_{20}^{|\lambda_1|} \; z_{10}^{|\lambda_2|}$$

for all but four orbits (besides the origin itself).

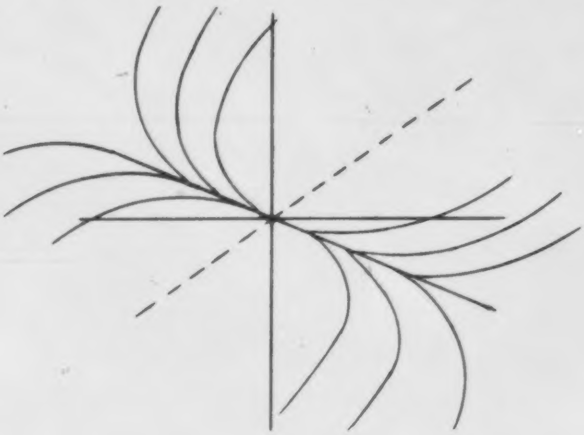
Question A.3,12. Explore these orbits; they look like hyperbolae, roughly speaking. Find the four exceptional orbits.

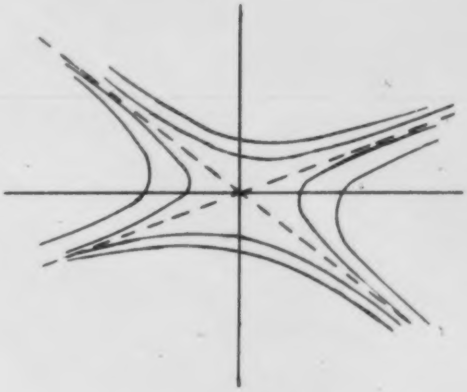
1°)D. $\lambda_2 < 0 < \lambda_1$. The reader will handle this case.

Question A.3,13. The reader will put arrows on the diagrams of the orbits in cases 1°)A, 1°)B, 1°)C, 1°)D.

The reader will check the following statement: in the above-discussed special cases the origin is an attractor if the real part of both eigenvalues is negative; the origin is a repulsor if the real part of both eigenvalues is positive. The examples we have seen where both real parts are zero, showed orbits that neither approached the origin nor receded from it. If one eigenvalue is negative and the other positive we have the peculiar arrangement of the saddle point: 2 orbits approach the origin; 2 orbits recede from it; the others first approach somewhat (if the initial point is located properly) and then recede.

Remark A.3,3. Terminology: The singular point is called an *improper node* if all orbits (with the exception at most of three) have the same asymptotic direction, either when $t \uparrow \infty$ or when $t \downarrow -\infty$. The singular point is called a *proper node* if the orbits are all straight lines with different directions. The singular point is called a *saddle point* when the two eigenvalues are real, one positive and one negative. The singular point is called a *spiral point* (or *vortex*, or *focus*) when the two eigenvalues are strictly complex (not pure imaginary). The singular point is called a *center* if all orbits are closed curves around it. The singular point $\vec{0}$ is an *attractor* if $\lim\limits_{t\uparrow+\infty} \vec{z}(t) = \vec{0}$, and a repulsor (or repeller) if the distance between it and $\vec{z}(t)$ increases as $t \uparrow +\infty$. In linear systems a repulsor becomes an attractor when time is reversed ($t = -\tau$). In order to avoid all ambiguity, we should point out that the asymptotic direction which is common to all (with the exception of at most three) orbits in the case of an *improper node*, is the direction of these orbits as they *approach* the *singular point*: we are talking about the tangent line to these orbits at the singular point after we have "added" the singular point to each orbit. In the case of a *saddle point* for a linear system the two lonely orbits that tend to the singular point as $t \uparrow +\infty$ and the two lonely orbits that tend to the singular point as $t \downarrow -\infty$ (going to infinity as $t \uparrow +\infty$) are called *separatrices*; the first two will be referred to as the *incoming separatrices*; the latter two, as the *outgoing separatrices*. The separatrices constitute the asymptotes of the other, hyperbola-like orbits.





Now we want to pick up the threads of the argument which we left when we introduced the distinction between six different classes of pairs of eigenvalues of the matrix A (a few pages into section A-3): after that point we discussed special representatives of the six classes of matrices A corresponding to these six classes of pairs of eigenvalues, but now we want to make the discussion completely general. We will use the same numbers, 1°) through 6°), to indicate the various cases.

1°) If the 2×2 matrix A has two different, non-zero, real eigenvalues, say λ_1 and λ_2, then we can construct a non-singular matrix V such that

$$A\,V = V\,\Lambda, \quad \text{or} \quad V^{-1}\,A\,V = \Lambda,$$

where Λ is the diagonal matrix with elements λ_1 and λ_2, and V^{-1} plays the role of the matrix P quoted in the paragraph on the 'basic strategy' below equation (A.3,1b).

Question A.3,14. Since $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, the matrix A is nonsingular (why?).

Hint: Your answer depends on your criterion for a nonsingular matrix: a matrix A is nonsingular iff there is no non-null vector $\vec{v}$ with $A\,\vec{v} = \vec{0}$, or iff det A $\neq 0$. With the earlier definition of eigenvalues as the roots of $\det(A - \lambda I) = 0$) the second criterion for A being nonsingular yields immediate results for our question (show!). Application of the first criterion needs the known result from linear algebra saying that for any given matrix A a non-null vector $\vec{v}$ and a (real or complex) number λ can satisfy the equation

$$A\,\vec{v} = \lambda\,\vec{v}$$

if and only if

$$(A - \lambda I)\,\vec{v} = \vec{0},$$

which is equivalent to

$$\det(A - \lambda I) = 0;$$

this shows that the λ in $A\,\vec{v} = \lambda\,\vec{v}$ is an eigenvalue. (Be it known that the $\vec{v}$ in $A\,\vec{v} = \lambda\,\vec{v}$ is called an eigenvector!) Now note that

$$A\,\vec{v} = \vec{0} \quad \text{is equivalent to} \quad A\,\vec{v} = 0\cdot\vec{v},$$

which shows that the existence of a non-null vector $\vec{v}$ with $A\vec{v} = 0$ is equivalent to the existence of an eigenvalue 0.

Question A.3,15. If $\vec{v}_1$ and $\vec{v}_2$ are both non-null vectors, then show that

$$c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$$

implies that either c_1 and c_2 are both zero, or both non-zero. (c_1, c_2 are real or complex numbers; a null vector has all components zero; a non-null vector has at least one non-zero component.)

For the further development we need the concept of linear (in)dependence of two vectors: two vectors are linearly dependent iff one is a scalar multiple of the other, $\vec{v}_2 = \gamma\vec{v}_1$ or $\vec{v}_1 = \delta\vec{v}_2$, which is equivalent to the definition that two vectors are linearly dependent iff a non-null vector $\vec{c}$ (with components c_1 and c_2) exists such that

$$V\vec{c} = \vec{0} \quad \text{or} \quad c_1\vec{v}_1 + c_2\vec{v}_2 = 0$$

(According to Question A.3,15 neither c_1 nor c_2 could be zero in this definition.) Obviously, the 2 × 2 matrix V built from the column vectors $\vec{v}_1$ and $\vec{v}_2$ is singular iff the vectors are dependent. Why? See preceding questions.

Question A.3,16. Now prove that if

$$\left\{\begin{array}{l} A\vec{v}_1 = \lambda_1\vec{v}_1 \\ \\ A\vec{v}_2 = \lambda_2\vec{v}_2 \end{array}\right\}, \quad \vec{v}_1 \neq \vec{0}, \quad \vec{v}_2 \neq \vec{0}, \quad \text{then}$$

$\vec{v}_1$ and $\vec{v}_2$ are linearly independent if $\lambda_1 \neq \lambda_2$.

Hint: Let $c_1\vec{v}_1 + c_2\vec{v}_2 = \vec{0}$. Then the equality $\vec{0} = (A - \lambda_1 I)(c_1\vec{v}_1 + c_2\vec{v}_2) = c_2(\lambda_2 - \lambda_1)\vec{v}_2$ entails that $c_2 = 0$, hence that $c_1 = 0$ (why?).

So, whenever a matrix A has two different eigenvalues, say λ_1 and λ_2, then any one eigenvector corresponding to the first eigenvalue and any one eigenvector corresponding to the second eigenvalue are linearly independent. If a 2 × 2 matrix A has the maximum number (i.e., 2) of different eigenvalues, then at least one eigenvector corresponds to each of these eigenvalues. Let $\vec{v}_1$ be an eigenvector for λ_1 and $\vec{v}_2$ for λ_2, then any other eigenvector (whether for λ_1 or for λ_2) must be of the form $\gamma_1\vec{v}_1 + \gamma_2\vec{v}_2$ (why?).

Question A.3,17. After $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$ have been found as the unequal roots of $\det(A - \lambda I) = 0$, $\vec{v}_i$ is uniquely (except for a scalar coefficient) determined by

$$A\,\vec{v}_i = \lambda_i\,\vec{v}_i, \quad i = 1, 2.$$

Prove! Also, the matrix

$$V = [\vec{v}_1 \quad \vec{v}_2]$$

is nonsingular (why?). As an exercise find the eigenvectors of

$$A' = \begin{bmatrix} 3 & -4 \\ 2 & 9 \end{bmatrix} \quad \text{and} \quad A'' = \begin{bmatrix} -1 & 3 \\ 2 & -2 \end{bmatrix}.$$

Also, for these two examples verify numerically the formula

$$V^{-1}\,A\,V = \Lambda,$$

quoted in the paragraph preceding Question A.3,14. Also verify that scalar coefficients do not matter: the equality $V^{-1}\,A\,V = \Lambda$ is just as valid, and with precisely the same λ_1 and λ_2, when we replace $\vec{v}_1$ by $q_1\vec{v}_1$ and $\vec{v}_2$ by $q_2\vec{v}_2$.

Question A.3,18. Now execute the program as set forth on the second page of section A.3. Transform the systems

$$\dot{\vec{x}} = A'\,\vec{x} \quad \text{and} \quad \dot{\vec{x}} = A''\,\vec{x},$$

where the single and double primes refer to the two matrices quoted in Question A.3,17, into systems

$$\dot{\vec{z}} = \Lambda'\,\vec{z} \quad \text{and} \quad \dot{\vec{z}} = \Lambda''\,\vec{z}$$

by means of the transformations

$$\vec{x} = V'\,\vec{z} \quad \text{and} \quad \vec{x} = V''\,\vec{z}.$$

Find all kinds of asymptotic directions in the $\vec{z}$-system and, consequently, in the $\vec{x}$-system.

Also write the solution functions both in $\vec{z}$-coordinates and in $\vec{x}$-coordinates.

The answers to questions A.3,14 through A.3,18 lead to the following summary of case 1°): If the 2×2 matrix A has two different, non-zero, real eigenvalues λ_1 and λ_2, then λ_1 determines an eigenvector $\vec{v}_1$ and λ_2 determines an eigenvector $\vec{v}_2$, both essentially unique and such that the matrix

$$V = [\vec{v}_1 \quad \vec{v}_2]$$

is nonsingular. The transformation

$$\vec{x} = V \vec{z}$$

transforms the system

$$\dot{\vec{x}} = A \vec{x}$$

into

$$\dot{\vec{z}} = V^{-1} A V \vec{z} = \Lambda \vec{z}.$$

The solutions of the z-system are given by

$$\vec{z}(t) = e^{(t-t_o)\Lambda} \vec{z}_o,$$

with transparent notation. The solutions of the x-system are given by

$$\vec{x}(t) = V \vec{z}(t) = V e^{(t-t_o)\Lambda} \vec{z}_o = V e^{(t-t_o)\Lambda} V^{-1} \vec{x}_o. \quad {}^{*)}$$

As for the asymptotic directions: in the case of an improper or proper node the preponderant asymptotic direction is that of the

$$\begin{cases} z_1\text{-axis} \quad (z_2 = 0), \quad \text{i.e. the vector } \vec{v}_1 \text{ iff } \lambda_2 < \lambda_1 < 0 \text{ or } 0 < \lambda_1 < \lambda_2 \\ z_2\text{-axis} \quad (z_1 = 0), \quad \text{i.e. the vector } \vec{v}_2 \text{ iff } \lambda_1 < \lambda_2 < 0 \text{ or } 0 < \lambda_2 < \lambda_1 \, ; \end{cases}$$

in the case of a saddle point the "incoming separatrix" is carried by the z_1-axis (i.e., by the line through the singular point in the direction of the vector $\vec{v}_1$) iff $\lambda_1 < 0 < \lambda_2$, and by the z_2-axis (i.e., by the similar line parallel to $\vec{v}_2$) iff $\lambda_2 < 0 < \lambda_1$.

Question A.3,19. Implement all details of the preceding summary and give a complete description of the solutions of

$$\dot{\vec{x}} = A' \vec{x} \quad \text{and} \quad \dot{\vec{x}} = A'' \vec{x}$$

with A' and A" as defined in Question A.3,17.

---

*) Again the origin, the unique singular point of $\dot{\vec{x}} = A \vec{x}$, is an attractor iff both λ_1 and λ_2 are negative.

2°) If the 2×2 matrix A has two eigenvalues that are real and equal, say λ_o, as well as non-zero, then there are two possibilities.

2°)A: Two linearly independent vectors $\vec{v}_1$ and $\vec{v}_2$ satisfy the equation $A\,\vec{v} = \lambda_o\vec{v}$ or $(A - \lambda_o I)\vec{v} = \vec{0}$. In this case a transformation is not even needed: linear algebra shows that now $(A - \lambda_o I)$ must be of rank 0. This is possible only if every element of $(A - \lambda_o I)$ is zero. So

$$A = \lambda_o I\,;$$

the general case coincides here with the special case.

2°)B: Only one vector $\vec{v}_1$ (except for a scalar coefficient) satisfies the equation $A\,\vec{v} = \lambda_o\vec{v}$ or $(A - \lambda_o I)\vec{v} = \vec{0}$. In this case the reader will show that he can take any vector $\vec{w}$, not a multiple of $\vec{v}_1$, and that then a real number γ exists so that

$$(A - \lambda_o I)\vec{w} = \gamma\,\vec{v}_i = \vec{u}, \quad \text{say.}$$

Then

$$A[\vec{u}\ \ \vec{w}] = [A\vec{u}\ \ A\vec{w}] = [\lambda_o\vec{u}\ \ \vec{u} + \lambda_o\vec{w}] = [\vec{u}\ \ \vec{w}]\begin{bmatrix}\lambda_o & 1\\ 0 & \lambda_o\end{bmatrix}.$$

Denoting $[\vec{u}\ \ \vec{w}]$ by V and

$$\begin{bmatrix}\lambda_o & 1\\ 0 & \lambda_o\end{bmatrix} \text{by } \Lambda,$$

we are back to our old tricks,

$$A\,V = V\,\Lambda.$$

The reader can complete the transformation of the present general case to the former special case 2°)B. The asymptotic direction of the orbits approaching the singular point is that of the z_1-axis, that of the vector $\vec{v}_1$.

Question A.3,20. Implement this procedure for $\dot{\vec{x}} = A\,\vec{x}$ with

$$A = \begin{bmatrix}-1 & 1\\ -9 & 5\end{bmatrix}.$$

Question A.3,21. In the above development the following results for 2×2 matrices A were used:

α) If $\det(A - \lambda I) = 0$ has two equal roots λ_o, then every element of the matrix $(A - \lambda_o I)^2$ is zero (a matter of patient, elementary algebra);

β) Defining $\vec{u} = (A - \lambda_o I)\vec{w}$, show that $(A - \lambda_o I)\vec{u} = \vec{0}$, which means that

$$\vec{u} = \gamma\, \vec{v}_1 \quad \text{(why?)}$$

γ) V is nonsingular provided $\vec{w}$ is chosen not to be a multiple of $\vec{v}_1$; prove!

3°) and 4°) If the 2×2 matrix A has two eigenvalues that are complex and different, say $\alpha + i\beta$ and $\alpha - i\beta$ with $\beta \neq 0$, and if $\vec{u}$ is the eigenvector corresponding to the eigenvalue $\alpha + i\beta$,

$$A\,\vec{u} = (\alpha + i\beta)\vec{u},$$

then it is obvious that the components of $\vec{u}$ cannot be real or pure imaginary (why not?). So put

$$\vec{u} = \vec{v} + i\vec{w}, \quad \text{with} \quad \vec{v} \neq \vec{0}, \quad \vec{w} \neq \vec{0},$$

where the components of $\vec{v}$ and $\vec{w}$ are real. Taking the real and the imaginary part of the resulting equality,

$$A(\vec{v} + i\vec{w}) = (\alpha + i\beta)(\vec{v} + i\vec{w}),$$

we find

$$A\,\vec{v} = \alpha\,\vec{v} - \beta\,\vec{w}$$

$$A\,\vec{w} = \beta\,\vec{v} + \alpha\,\vec{w}$$

> Question A.3,22. Prove that $\vec{v}$ and $\vec{w}$ are linearly independent. Because of the result of Question A.3,15 we only need to prove that $c_1\vec{v} + c_2\vec{w} = \vec{0}$ entails that c_1 and c_2 cannot be both different from zero: the only remaining possibility is then that both are zero, which proves our claim. Suppose c_1 and c_2 are both nonzero. Then
>
> $$c_1\vec{v} = -c_2\vec{w}.$$
>
> Then show that the above equality $A(\vec{v} + i\vec{w}) = (\alpha + i\beta)(\vec{v} + i\vec{w})$ yields
>
> $$A\,\vec{w} = (\alpha + i\beta)\vec{w},$$
>
> which contradicts $\beta \neq 0$ (why?). So the assumption that $c_1 \neq 0$, $c_2 \neq 0$, leads to a contradiction. Hence $c_1 = 0 = c_2$; and $\vec{v}$ and $\vec{w}$ are linearly independent.

The pair of equations preceding Question A.3,22 can be written as

$$A\ V = V\ \Lambda$$

where $V = [\vec{v}\ \ \vec{w}]$ is nonsingular and $\Lambda = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}$. According to our usual method the solutions of

$$\dot{\vec{x}} = A\ \vec{x}$$

are of the form

$$\vec{x}(t) = V\ \vec{z}(t),$$

where

$$\vec{z}(t) = e^{(t-t_o)\alpha} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \vec{z}(t_o)$$

with $c = \cos \beta(t-t_o)$, $s = \sin \beta(t-t_o)$, $\vec{z}(t_o) = V^{-1}\ \vec{x}(t_o)$. Thus $\vec{z}(t)$ satisfies the equality

$$\vec{z}^{\,T}(t)\ \vec{z}(t) = e^{2(t-t_o)\alpha}\ \vec{z}^{\,T}(t_o)\vec{z}(t_o),$$

from which it follows that if $\alpha \overset{<}{\underset{>}{=}} 0$, the point $\vec{z}(t)$, describing a solution curve, finds itself on

$$\begin{Bmatrix} \text{gradually smaller} \\ \text{one and the same} \\ \text{gradually bigger} \end{Bmatrix}$$

circle perimeter(s). The origin, a unique singular point for this system, is in these three respective cases

$$\begin{Bmatrix} \text{attractor} \\ \text{center} \\ \text{repulsor} \end{Bmatrix}$$

Similarly, in terms of $\vec{x}(t)$:

$$\vec{x}^{\,T}(t)\ (V\ V^T)^{-1}\ \vec{x}(t) = e^{2(t-t_o)\alpha}\ \vec{x}_o^{\,T}\ (V\ V^T)^{-1}\ \vec{x}_o.$$

So, if $\alpha \overset{<}{\underset{>}{=}} 0$, the point $\vec{x}(t)$, describing a solution curve, finds itself on

$$\begin{Bmatrix} \text{gradually smaller} \\ \text{one and the same} \\ \text{gradually bigger} \end{Bmatrix}$$

ellipse perimeter(s). Again the origin (singular point) is an attractor iff $\alpha < 0$.

5°) One eigenvalue zero. The equation for the eigenvalues, $\det(A - \lambda I) = 0$, can be written as $\lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a22 - a_{12}a_{21} = 0$. One root is zero iff $a_{11}a_{22} = a_{12}a_{21}$, which amounts to

α) $a_{11}/a_{21} = a_{21}/a_{22} = p^{-1}$, say; putting $a_{11} = a$ and $a_{12} = b$, we find

$$A = \begin{bmatrix} a & b \\ pa & pb \end{bmatrix}.$$

β) $a_{12} = 0 = a_{22}$; special case of α) (elaborate!)

γ) $a_{22} = 0 = a_{21}$; special case of α) (elaborate!)

δ) other choices of zero elements reduce to (β) or (γ) by reassignment of subscripts.

Note that the other eigenvalue is nonzero iff in case (α) $\underline{a + pb \neq 0}$, in cases (β) and (γ) $\underline{a_{11} \neq 0}$.

Question A.3,23. Since cases (β), (γ), (δ) are all special cases of (α), we need to study only (α). The reader will prove that the nonzero eigenvalue is $a + pb$; its eigenvector has components 1 and p; the other eigenvector has components $-b$ and a; these two eigenvectors are linearly independent; finally

$$A\ V = V\ \Lambda$$

with A as defined under point (α) above, and with

$$V = \begin{bmatrix} 1 & -b \\ p & a \end{bmatrix}, \quad \Lambda = \begin{bmatrix} a + pb & 0 \\ 0 & 0 \end{bmatrix}.$$

The results of Question A.3,23 show how our general case (5°) reduces to the earlier-discussed special case of (5°), where

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}$$

No singular point is an attractor in this system!

6°) Both eigenvalues zero. In the matrix A quoted under the above point (α) in (5°) we must now put $a + pb = 0$ (why?). This reduces matrix A to the special form

$$A = \begin{bmatrix} -pb & b \\ -p^2 b & pb \end{bmatrix}.$$

We have to distinguish two cases here:

6°)A. b = 0. Then A is of the same form as the earlier-discussed special case 6°A; all elements of A equal to zero.

6°)B. $b \neq 0$, p arbitrary. The reader will show that now

$$A\ V = V\ \Lambda,$$

with

$$A = \begin{bmatrix} -pb & b \\ -p^2 b & pb \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 0 \\ p & b^{-1} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix};$$

Note that V is nonsingular. The reader will find all singular points and verify that none of them is an attractor.

Remark A.3,4. This concerns finding the asymptotic directions for nodes, the directions for the separatrices for saddle points, and the absence of asymptotic directions for spiral points. In Questions A.3,9, A.3,10, and A.3,11, in the discussion preceding Question A.3,19, and in the last line before Question A.3,20 we made observations regarding asymptotic directions of orbits in the neighborhood of an improper node: the preponderant asymptotic direction is given by that eigenvector of the system's matrix A which corresponds to the eigenvalue which is closer to zero, and the exceptional asymptotic direction is given by the eigenvector corresponding to the other eigenvalue. Similarly, in Question A.3,12, in the latter part of Remark A.3,3, and in the discussion preceding Question A.3,19 we showed that the direction of the incoming separatrix around a saddle point is given by that eigenvector of the system's matrix A which corresponds to the negative eigenvalue, whereas the direction of the outgoing separatrix is given by the eigenvector corresponding to the positive eigenvalue. Now there is a fast method to obtain the slopes of these two important directions, which, however, drops the distinction between the two;

but in many cases, when you construct a diagram of the phase portrait and sketch the two directions, this distinction is evident from the quick-and-dirty isocline and arrow method. Thus, let k_1 and k_2 be the slopes of the (preponderant and the exceptional) asymptotic directions around an improper node, or let k_1 and k_2 be the slopes of the (incoming and outgoing) separatrices around a saddle point. Then k_1 and k_2 are the roots of the quadratic equation

$$a_{12}k^2 + (a_{11} - a_{22})k - a_{21} = 0,$$

where

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = A$$

is the matrix of the system $\dot{\vec{x}} = A\vec{x}$. Note that there are no real roots for k when

$$(a_{11} - a_{22})^2 + 4\, a_{12}a_{21} < 0.$$

The reader will have no difficulty in showing that the eigenvalues of A are non-real under the same condition. So spiral points can be identified on the basis of nothing but the just-introduced k-equation. The distinction between a node and a saddle point follows geometrically, from the isocline-and-arrow method, or analytically, from the signs of the eigenvalues of the matrix A. The distinction between stable and unstable nodes or spiral points follows from the signs of the real parts of the eigenvalues of A. The distinction between stable and unstable nodes also follows from the arrow method, but not the distinction between stable and unstable spiral points.

Question A.3,24. Let λ_i be an eigenvalue of the 2×2 matrix A, and $\vec{v}_i$ be the corresponding eigenvector, with the components v_{1i} and v_{2i} $(i = 1, 2)$. Then the reader will show that

$$k_i = \frac{v_{2i}}{v_{1i}} = \frac{\lambda_i - a_{11}}{a_{12}} \qquad (i = 1, 2)$$

$$k_1 + k_2 = \frac{\lambda_1 + \lambda_2 - 2a_{11}}{a_{12}} = \frac{a_{22} - a_{11}}{a_{12}}$$

$$k_1 k_2 = \frac{-a_{21}}{a_{12}}$$

which directly shows that k_1 and k_2 are the roots of

$$a_{12}k^2 + (a_{11} - a_{22})k - a_{21} = 0.$$

Elaborate this proof!

Remark A.3,5. What if in $\dot{\vec{x}} = A\vec{x}$ the vector $\vec{x}$ is an n-vector and the matrix A is an $n \times n$ matrix, $n > 2$? There is then a much greater variety of possibilities, more or less completely recorded only for $n = 3$ (e.g., see J. W. Reyn, 1964, Classification and description of the singular points of a system of three differential equations; Z. Angew. Math. Phys. 15, pp. 540-557). One result that carries over rather straightforward is the one about stability: when all eigenvalues λ of the matrix A have strictly negative real parts, then all solution curves approach the singular point $\vec{0}$ as $t \uparrow +\infty$. When all $\operatorname{Re} \lambda > 0$, then all solutions $\to \vec{0}$ as $t \downarrow -\infty$. Whenever some eigenvalues have positive real parts and others have negative real parts, the situation is as in the above saddle points: a great many trajectories approach $\vec{0}$ to a degree, and then take off for infinity. We will not go into the finer points of n-dimensional phase portraits.

A.4. Linearization.

What is called linearization consists of the following method of investigating the phase protrait locally, i.e., in the neighborhood of any singular point. Given the nonlinear autonomous differential system

$$\dot{\vec{x}}(t) = \vec{f}(\vec{x}(t)) \qquad (A.4,1)$$

with a singular point $\vec{\alpha}$,

$$\vec{f}(\vec{\alpha}) = \vec{0},$$

$\vec{f}$ mapping n-vector space into n-vector space, perform the following steps:

1) Transform $\vec{x}(t) = \vec{\alpha} + \vec{y}(t)$:

$$\dot{\vec{y}}(t) = \vec{f}(\vec{\alpha} + \vec{y}(t)). \tag{A.4,2}$$

2) Develop the second member of (A.4,2) in a truncated Taylor series,

$$\vec{f}(\vec{\alpha} + \vec{y}) = \vec{f}(\vec{\alpha}) + D\vec{f}(\vec{\alpha})\cdot\vec{y} + \sigma(\|\vec{y}\|),$$

where $D\vec{f}(\vec{\alpha})$ is the matrix with (i,j)-element $D_j f_i(\vec{\alpha})$, that is, the value of the partial derivative of the i^{th} component of the vector $\vec{f}(\vec{x})$ with respect to the j^{th} variable x_j, evaluated at $\vec{x} = \vec{\alpha}$. Note that $\vec{f}(\vec{\alpha}) = \vec{0}$.

3) Study the linear system

$$\dot{\vec{y}}(t) = D\vec{f}(\vec{\alpha})\,\vec{y}(t), \tag{A.4,3}$$

especially the behavior of its solutions in the neighborhood of $\vec{y} = \vec{0}$.

4) Assert that in a neighborhood of $\vec{\alpha}$ the (asymptotic) behavior of the trajectories of $\dot{\vec{x}} = \vec{f}(\vec{x})$ is "like" that of the trajectories of $\dot{\vec{y}} = D\vec{f}(\vec{\alpha})\,\vec{y}$ in a neighborhood of $\vec{0}$.

Essentially then, the linearization method substitutes the study of (one or more) linear systems for the study of a nonlinear system in the neighborhood of its (one or more) singular points. The local results are valid only under certain conditions (concerning the eigenvalues of $D\vec{f}(\vec{\alpha})$ as well as the nature of the correction term $\sigma(\|\vec{y}\|)$): and the word "like" in the above point (4) has to be clarified and specified. The proofs concerning this method are too complicated for our present purposes. We'll first give a few examples and counterexamples, and then list some of the theorems. For the routine applications of the linearization method one does not do much else than investigate the λ-equation and the k-equation connected with the matrix $D\vec{f}(\vec{\alpha})$.

Question A.4,1. The differential equation $\dot{x} = x(1 - x)$ has two singular points, $x = 0$ and $x = 1$. By the linearization method show that $x = 0$ is unstable and $x = 1$ is stable.

Hint: Since this equation is 1-dimensional we can execute the whole program without pain:
Find the Taylor series around $x = 0$, and the Taylor series around $x = 1$. Find the substitute linear differential equations in y:

$$\begin{array}{ccc} \dot{\vec{y}} = y & \qquad & \dot{\vec{y}} = -y \\ \Downarrow & & \Downarrow \\ y(t) = e^{t-t_o} y_o & & y(t) = e^{-(t-t_o)} y_o \end{array}$$

The conclusions re stability and instability follow.

Alternatively, $\vec{Df}(\vec{\alpha})$ is just a real number here instead of a matrix, or equivalently, $Df(\alpha)$ is a 1×1 matrix; its eigenvalue equals $Df(\alpha)$. For $f(x) = x - x^2$, find $Df(x)$, $Df(0)$, $Df(1)$ and show your conclusions as based on these eigenvalues are the same as those following from the first approach, where you actually solved the substitute linear systems.

Question A.4,2. The system

$$\left\{ \begin{aligned} \dot{x}_1 &= x_1(\alpha_1 - b_{11} \log x_1 - b_{12} \log x_2) \\ \dot{x}_2 &= x_2(\alpha_2 - b_{21} \log x_1 - b_{22} \log x_2) \end{aligned} \right\} \qquad (A.4,4)$$

has (among others) a singular point at (ξ_1, ξ_2) as determined by

$$\left\{ \begin{aligned} b_{11} \log \xi_1 + b_{12} \log \xi_2 &= \alpha_1 \\ b_{21} \log \xi_1 + b_{22} \log \xi_2 &= \alpha_2 \end{aligned} \right\}$$

Find $\vec{Df}(\vec{\xi})$ and show that the corresponding λ-equation is

$$\lambda^2 + (b_{11} + b_{22})\lambda + (b_{11}b_{22} - b_{12}b_{21}) = 0.$$

Also make the substitution $\log x_i = y_i$ in equation (A.4,4), and in the resulting equations for y_i make the substitution $y_i = \eta_i + z_i$, where $\eta_i = \log \xi_i$. The singular point $\vec{z} = \vec{0}$ corresponds to $\vec{x} = \vec{\xi}$. The differential equation for $\vec{z}(t)$ can be solved explicitly and the solutions for $\vec{x}(t)$ can be expressed in terms of the $\vec{z}$-solutions. The reader may check that $\vec{\xi}$ is stable iff $b_{11}b_{22} - b_{12}b_{21} > 0$, $b_{11} + b_{22} > 0$: a) directly from the explicit solutions, b) from the above λ-equation. The reader may also check that $\vec{\xi}$ is a stable singular point for the $\vec{x}$-equation iff $\vec{0}$ is a stable singular point for the $\vec{z}$-equation. Note that the amount of algebra needed to find the λ-equation from $\vec{Df}(\vec{\xi})$ is negligible relative to finding the explicit form of the x-solutions. If the reader is a glutton for work, he may want to put $\vec{x}(t) = \vec{\xi} + \vec{e}(t)$, find the differential system for $\vec{e}(t)$, develop the necessary Taylor series and find the linear parts in $e_1(t)$ and $e_2(t)$. The result should be easy to explain in the light of the above.

Question A.4,3. The system

$$\left\{\begin{array}{l} \dot{x}_1 = -\lambda_1 x_1 + x_2^2 \\ \\ \dot{x}_2 = x_1^2 - \lambda_2 x_2 \end{array}\right\} \qquad (A.4,5)$$

has two singular points, $(0, 0)$ and $\left((\lambda_1 \lambda_2^2)^{1/3}, (\lambda_1^2 \lambda_2)^{1/3}\right)$. Analyze $\vec{Df}(\vec{x})$ at both points. Find the eigenvalues and the slopes of the important directions. If $0 < \lambda_1 < \lambda_2$, then the x_1-axis is the preponderant asymptotic direction around the origin. Sketch some trajectories in that area, in detail (as if under a miscroscope).

Question A.4,4.

$$\left.\begin{array}{l} \left.\begin{array}{l} \dot{x}_1 = -x_1 - \dfrac{x_2}{\log\sqrt{x_1^2 + x_2^2}} \\ \\ \dot{x}_2 = -x_2 + \dfrac{x_1}{\log\sqrt{x_1^2 + x_2^2}} \end{array}\right\} \text{ for } x \neq 0 \\ \\ \dot{\vec{x}} = \vec{0} \qquad\qquad\qquad\qquad \text{ for } x = 0. \end{array}\right\} \qquad (A.4,6)$$

The solutions of this system can be constructed explicitly for $x_1^2 + x_2^2 < 1 - \delta$, where $0 < \delta < 1$, and they come out quite different from the solutions of the linearized system (cf. Question A.4,6),

$$\left\{\begin{array}{l} \dot{x}_1 = -x_1 \\ \\ \dot{x}_2 = -x_2 \end{array}\right\} \qquad (A.4,6a)$$

Whereas the solutions of (A.4,6a) are $\vec{x}(t) = e^{-(t-t_0)}\vec{x}_0$, with trajectories in the phase plane,

$$x_2(t)/x_1(t) = x_{20}/x_{10}$$

(straight lines radiating from the singular point at the origin), the solutions of (A.4,6) are found after introduction of polar coordinates: $x_1(t) = r(t) \cos \varphi(t)$, $x_2(t) = r(t) \sin \varphi(t)$. Equations (A.4,6) then reduce to

$$\left\{\begin{aligned} \dot{r}(t) &= - r(t) \\ \dot{\varphi}(t) &= \frac{1}{\log r(t)} \end{aligned}\right\} \tag{A.4,6b}$$

with solutions defined by

$$\left\{\begin{aligned} r(t) &= e^{-(t-t_o)} r_o \\ \varphi(t) &= - \log(t - t_o - \log r_o) \end{aligned}\right\} \tag{A.4,7}$$

Note that we are studying equation (A.4,6) only for $x_1^2 + x_2^2 < 1 - \delta$; so $t - t_o - \log r_o > t - t_o \geqq 0$, for all time later than t_o. The most obvious feature of these solutions is that they are spirals, winding around the origin, differing indeed from the straight lines of (A.4,6a).

Question A.4,5. Although there is a great deal of difference between the straight-line solutions of equation (A.4,6a) and the spiral solutions (see Equation (A.4,7)) of equation (A.4,6), the two do agree in one important respect: if they start at a point in the interior of the unit circle about the origin, then they tend to the origin in the limit, as $t \uparrow +\infty$. Check this! Note that the λ-equation for (A.4,6a) has non-zero roots. The reason for the discrepancy between the phase-portraits of (A.4,6) and of (A.4,6a) is that the remainder terms, $x_i/\log r$, where $r = \sqrt{x_1^2 + x_2^2}$, do not tend to zero fast enough (see the relevant theorems below).

Question A.4,6. In Question A.4,4 it was asserted that (A.4,6a) is the linearized system of (A.4,6). That looks very plausible indeed, since the terms $x_i/\log r$ are nonlinear and tend to zero faster than the linear terms. However, the description given of the linearization method in the beginning of section A.4, requires that we determine the matrix $D\vec{f}(\vec{0})$ (since the singular point of (A.4,6) is the origin, cf. Question A.4,7).

The reader will now actually compute the elements of the matrix $D\vec{f}(\vec{0})$.

Hint: It is vastly more efficient to compute these partial derivatives at the origin directly from their definition, such as

$$\lim_{h \to 0} \frac{f_1(h,0) - f_1(0,0)}{h},$$

than from their general expressions for arbitrary x_1 and x_2.

Question A.4,7. Prove that the origin is the only singular point of the system defined by equation (A.4,6). Hint: Assume it is not; then multiply the first equation by x_1, the second by x_2, and add. You find that the singular point must satisfy the equation $x_1^2 + x_2^2 = 0$.

Whereas the few preceding examples illustrate the fact that stability and instability are the same between a nonlinear system and its linearization, even in the fact of marked differences between the phase portraits, provided the λ-values have non-zero real parts, the following few examples will illustrate the fact that practically all bets are off if the λ-values have zero real parts. In such cases the linearized system may for instance show a continuum of singular points in any neighborhood of the unique singular point of the nonlinear system, thereby disrupting any similarity in asymptotic behavior between the two systems.

Question A.4,8. Study the system

$$\left\{ \begin{array}{rcl} \dot{x}_1 & = & x_1^3 \\ \dot{x}_2 & = & -x_2 \end{array} \right\} \tag{A.4,8}$$

and its linearization re its singular point,

$$\left\{ \begin{array}{rcl} \dot{x}_1 & = & 0 \\ \dot{x}_2 & = & -x_2 \end{array} \right\} \tag{A.4,8a}$$

The phase portrait of this linear system is a collection of half-lines parallel to the x_2-axis. The phase portrait of the nonlinear system bears a strong resemblance to that of a linear system with a saddle point at the origin. The reader will fill in the details. Hint: The trajectories of (A.4,8) have the equation

$$x_2 = e^{\frac{1}{2x_1^2}} \; e^{-\frac{1}{2x_{10}^2}} \; x_{20} \, ;$$

note that these trajectories have different horizontal asymptotes.

Question A.4,9. Study the system

$$\left\{ \begin{aligned} \dot{x}_1 &= -x_1^3 \\ \dot{x}_2 &= -x_2 \end{aligned} \right\} \tag{A.4,9}$$

and its linearization re its singular point,

$$\left\{ \begin{aligned} \dot{x}_1 &= 0 \\ \dot{x}_2 &= -x_2 \end{aligned} \right\}. \tag{A.4,9a}$$

This linearized system is the same as in the preceding question, but the behavior of the nonlinear system is vastly different. The origin is now asymptotically stable in the large (i.e., wherever the initial point $\vec{x}_o$ is the corresponding solution will tend to the origin in the limit as $t \uparrow +\infty$), and looks pretty much like a stable node, with principal asymptotic direction that of the x_1-axis, and with the x_2-axis as carrier of two trajectories. Hint: The trajectories are now determined by

$$x_2 = e^{-\frac{1}{2x_1^2}} \; e^{+\frac{1}{2x_{10}^2}} \; x_{20}$$

As $t \uparrow +\infty$, $x_2(t)/x_{20} \downarrow 0$ and $|x_1(t)| \downarrow 0$. Show that for every trajectory (except the two carried by the x_2-axis, and the origin itself):

$$\frac{x_2}{x_1} \downarrow 0 \quad \text{as} \quad |x_1| \downarrow 0,$$

which shows that all these trajectories have the x_1-axis as asymptotic tangent line. What happens as $|x_1| \uparrow +\infty$? Sketch the trajectories!

Question A.4,10. Study the system

$$\left\{ \begin{aligned} \dot{x}_1 &= x_1^2 \\ \dot{x}_2 &= -x_2 \end{aligned} \right\} \tag{A.4,10}$$

One half of the phase portrait looks like that of (A.4,8); the other half, like that of (A.4,9). Elaborate!

Question A.4,11. Study the system

$$\left\{\begin{matrix} \dot{x}_1 = -x_2 + x_1(1 - r^2)r \\ \\ x_2 = \quad + x_2(1 - r^2)r \end{matrix}\right\}, \qquad (A.4,11)$$

where $r = \sqrt{x_1^2 + x_2^2} \geqq 0$. Change to polar coordinates:

$x_1 = r \cos \varphi$, $x_2 = r \sin \varphi$, and find

$$\left\{\begin{matrix} \dot{r}_1 = (1 - r^2)r^2 \\ \\ \dot{\varphi} \ = 1 \end{matrix}\right\} (r \geqq 0 \quad , \qquad (A.4,11a)$$

The first equation can be written as

$$\frac{\dot{r}}{r^2} + \tfrac{1}{2}\frac{\dot{r}}{1 - r} + \tfrac{1}{2}\frac{\dot{r}}{1 + r} = 1$$

which can be integrated to

$$e^{t-t_o} = e^{-\frac{1}{r} + \frac{1}{r_o}} \cdot \sqrt{\frac{1 + r}{1 + r_0} \cdot \frac{1 - r_o}{1 - r}}, \quad r_o > 0.$$

This equation shows (why?) that when $0 < r_o < 1$, then $r = r(t)$ will for all $t > t_o$ remain in the open interval $]0,1[$. For $0 < r < 1$ the second member of this equation is a monotone increasing function of r. So this equation does define r as a function of t. Prove that as $t \uparrow +\infty$, $r = r(t) \uparrow 1$. Noting that $\varphi(t) = t - t_o$, we find that the solutions of (A.4,11) consist of outward spirals toward the unit circle (given $0 < r_o < 1$): the origin is an unstable spiral point. The linearization of (A.4,11),

$$\left\{\begin{matrix} x_1 = - x_2 \\ \\ x_2 = + x_1 \end{matrix}\right\}, \qquad (A.4,11b)$$

has, however, a phase portrait consisting entirely of circles around the origin, which therefore is a center.

Among the many theorems re linearization, which are discussed in great detail in Andronov et al. (1973) (as well as in E. A. Coddington and N. Levinson, 1955, Theory of Ordinary Differential Equations; McGraw-Hill, xii + 429 pp.; Chap. 15; or in L. Cesari, 1963, Asymptotic Behavior and Stability Problems in Ordinary Differential Equations; Springer, viii + 271 pp.; sec. 9.[illegible]), we mention only a few. All refer to the situation in which the singular point under scrutiny is the origin (because of the transformation from equation (A.4,1) to equation (A.4,2) this does not hurt generality). All refer to the case in which the matrix $D\vec{f}(\vec{0})$ is non-singular, in fact to the case in which the eigenvalues of this matrix have non-zero real parts. Furthermore, certain restrictions need to be imposed on the nonlinear part of $\vec{f}(\vec{x})$, that is, on $\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x}$: even in the weaker theorem(s), viz., those on stability rather than on the shape of the phase portrait. First we'll discuss (local) stability, then mode of approach to the singular point.

Theorem A.4,1. Consider the differential equation

$$\dot{\vec{x}} = \vec{f}(\vec{x}). \tag{A.4,12}$$

Let $\vec{f}(\vec{0}) = \vec{0}$. Let all eigenvalues of the matrix $\underline{D\vec{f}(\vec{0})}$ have negative real part. Let

$$\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x} = \underline{o(\|\vec{x}\|)}. \tag{A.4,13}$$

Then the origin is an attractor for the equation $\dot{\vec{x}} = \vec{f}(\vec{x})$.
(For an example see Question A.4,12 below.)

Terminology. The meaning of equality (A.4,13) is this:

$$\lim_{\vec{x}\to\vec{0}} \frac{\|\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x}\|}{\|\vec{x}\|} = 0.$$

Here the vertical bars denote 'norm of.' The three most-used versions of norms will be

$$\|\vec{y}\| = \sqrt{y_1^2 + y_2^2}, \quad \|\vec{y}\| = |y_1| + |y_2|, \quad \text{and} \quad \|\vec{y}\| = \mathrm{Max}(|y_1|, |y_2|).$$

Terminology: The phrase, "the origin is an attractor for the equation $\dot{\vec{x}} = \vec{f}(\vec{x})$," needs to be defined somewhat more carefully for nonlinear equations than for linear ones. It is used here to mean the following: there is a real number $\epsilon > 0$ such that each solution $\vec{x}(t)$ of our equation that at some time τ satisfies

$$\|\vec{x}(\tau)\| < \varepsilon,$$

exists for all $t > \tau$ and has the property that

$$\lim_{t\uparrow\infty} \|\vec{x}(t)\| = 0.$$

Remark A.4,1. Note that $\|\vec{y}\| < \alpha$ has slightly different meanings, varying with the definition of norm in use at the moment. The reader will not find it difficult to show that according to the first definition

$$\|\vec{y}\| < \alpha \text{ implies } |y_1| < \alpha \text{ and } |y_2| < \alpha; \quad |y_1| < \alpha/\sqrt{2} \text{ and } |y_2| < \alpha/\sqrt{2} \text{ imply } \|y\| < \alpha.$$

Similarly according to the second definition

$$\|\vec{y}\| < \alpha \text{ implies } |y_1| < \alpha \text{ and } |y_2| < \alpha; \quad |y_1| < \alpha/2 \text{ and } |y_2| < \alpha/2 \text{ imply } \|\vec{y}\| < \alpha.$$

Similarly according to the third definition

$$\|\vec{y}\| < \alpha \text{ implies } |y_1| < \alpha \text{ and } |y_2| < \alpha; \quad |y_1| < \alpha \text{ and } |y_2| < \alpha \text{ imply } \|\vec{y}\| < \alpha.$$

Consequently, according to any of those definitions

$$\text{"}\lim_{t\uparrow\infty} \|\vec{x}(t)\| = 0\text{"} \quad \text{iff} \quad \text{"}\lim_{t\uparrow\infty} x_1(t) = 0 \text{ and } \lim_{t\uparrow\infty} x_2(t) = 0\text{"}$$

Question A.4,12. Consider the equation

$$\left\{\begin{aligned} \dot{x}_1 &= a_{11}x_1 + a_{12}x_2 + b_{11}x_1^2 + 2\,b_{12}x_1x_2 + b_{22}x_2^2 \\ \dot{x}_2 &= a_{21}x_1 + a_{22}x_2 + c_{11}x_1^2 + 2\,c_{12}x_1x_2 + c_{22}x_2^2 \end{aligned}\right\}. \qquad (A.4,14)$$

Prove that this equation satisfies (A.4,13), no matter what definition of norm is used. Hint: Show that what you want to prove is equivalent to

$$\lim_{\vec{x}\to\vec{0}} = \frac{b_{11}x_1^2 + 2\,b_{12}x_1x_2 + b_{22}x_2^2}{\sqrt{x_1^2 + x_1^2}} = 0,$$ and similarly with c instead

of b; then switch to polar coordinates. Therefore, as soon as in equation (A.4,14) the matrix A, with elements a_{ij}, has eigenvalues with negative real parts, the origin is an attractor.

Note that many of the examples in the main text of this module present such polynomials as occur in (A.4,14). So the version of Theorem A.4,1 that is used most often is the version of Question A.4,12: all you need to do with these polynomial models is to check the eigenvalues of the matrix of the linear part!

> Question A.4,13. Show that Theorem A.4,1 also takes care of equation (A.4,6) in Question A.4,4. It proves that the origin is an attractor for equation (A.4,6).
>
> Question A.4,14. Theorem A.4,1 can not take care of the differential equations discussed in Questions A.4,8, A.4,9 and A.4,10. Why not?

Theorem A.4,2. Consider the differential equation

$$\dot{\vec{x}} = \vec{f}(\vec{x}),$$

with $\vec{f}(\vec{0}) = \vec{0}$. If the origin is a spiral point for the equation

$$\dot{\vec{x}} = D\vec{f}(\vec{0})\cdot\vec{x}$$

and if

$$\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x} = o(\|\vec{x}\|),$$

then the origin is also a spiral point for

$$\dot{\vec{x}} = \vec{f}(\vec{x}).$$

Terminology. The phrase 'the origin is a spiral point for the equation $\dot{\vec{x}} = \vec{f}(\vec{x})$' needs to be defined more carefully for nonlinear equations than for linear ones. It is used here to mean either one of two things: a stable spiral point, or an unstable spiral point. The origin is a stable spiral point iff 1°) the origin is an attractor (see above; mind the role of $t \uparrow \infty$ in that definition) and 2°) in polar coordinates the solutions

$$\left\{\begin{aligned} x_1(t) &= r(t)\cos\varphi(t) \\ x_2(t) &= r(t)\sin\varphi(t) \end{aligned}\right\}$$

have the properties that $r(t) \to 0$ as $t \uparrow \infty$ (which tells nothing but that the origin is an attractor) and $|\varphi(t)| \uparrow \infty$ (this is what makes $\vec{0}$ a spiral point). The definition of the origin being an unstable spiral point is obtained from that of a stable spiral point by time reversal: put $t = -T$ and write all definitions with T instead of t.

Remark A.4,2. Theorem A.4,2 is somewhat remarkable in that it makes a statement about the mode of approach of the solution curves to the singular point, not just about stability or instability, under a rather weak condition, viz.

$$\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x} = \sigma(\|\vec{x}\|).$$

The example discussed in Question A.4,4 shows that this condition is not enough to keep a singular point a proper node under the transition from linearity to nonlinearity, but the spiral point property is apparently a bit sturdier. (For completeness' sake it should be mentioned that Coddington and Levinson, 1955, p. 378 (see paragraph following Question A.4,11) point out that under this weak condition a spiral point may cease to be 'proper' under the transition from linearity to nonlinearity, but we will not discuss this fine point).

Question A.4,15. The behavior of centers is quite the opposite from that of spiral points. If the origin is a center for

$$\dot{\vec{x}} = D\vec{f}(\vec{0})\cdot\vec{x},$$

then the origin may be a center, or a stable spiral point, or an unstable spiral point for

$$\dot{\vec{x}} = \vec{f}(\vec{x}),$$

even when the nonlinear part of $\vec{f}(\vec{x})$ satisfies more stringent conditions than (A.4,13).

The reader will check this for the following three equations

$$\left\{\begin{array}{l} \dot{x}_1 = x_2 + x_2 r \\ \dot{x}_2 = -x_1 - x_1 r \end{array}\right\}, \quad \left\{\begin{array}{l} \dot{x}_1 = x_2 - x_1 r \\ \dot{x}_2 = -x_1 - x_2 r \end{array}\right\}, \quad \left\{\begin{array}{l} \dot{x}_1 = x_2 + x_1 r \\ \dot{x}_2 = -x_1 + x_2 r \end{array}\right\}$$

where $r = \sqrt{x_1^2 + x_2^2}$. Another example of a nonlinear equation having a center at the same time as its linearization is given by the classical Volterra predator-prey-model (see Question 2.1,3).

Theorem A.4,3. Consider the differential equation ($\vec{x}$ being a 2-vector)

$$\dot{\vec{x}} = \vec{f}(\vec{x}),$$

with $\vec{f}(\vec{0}) = \vec{0}$. Let

$$\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x} = \sigma(\|\vec{x}\|).$$

Let the eigenvalues of $D\vec{f}(\vec{0})$ be real, strictly negative, and unequal, say $\lambda_2 < \lambda_1 < 0$. Then as $t \uparrow \infty$, every orbit of $\dot{\vec{x}} = \vec{f}(\vec{x})$ tends to the singular point $\vec{0}$ in one of four specific directions, an infinite number of them in the directions of $\vec{v}_1$ and $-\vec{v}_1$, and at least one of them in each of the directions of $\vec{v}_2$ and $-\vec{v}_2$. Here $\vec{v}_1$ is the eigenvector of $D\vec{f}(\vec{0})$ corresponding to λ_1, and $\vec{v}_2$ is the eigenvector of $D\vec{f}(\vec{0})$ corresponding to λ_2.

Under stronger conditions re the nonlinear part of $\vec{f}(\vec{x})$ the local phase portrait of our nonlinear equation is exactly like that of its linearization: one orbit in the direction of $\vec{v}_2$, and one in the direction of $-\vec{v}_2$. An example of such stronger conditions is the continuity of the derivatives of $\vec{f}(\vec{x})$ in a neighborhood of the singular point:

$$D_j f_i(x_1,x_2) \text{ continuous in a nbhd of } \vec{0}.$$

In accordance with p.126 of Andronov et al. (1973) this allows us to write the components of $\vec{f}(\vec{x})$ in the form

$$\left\{\begin{aligned} f_1(x_1,x_2) &= a_{11}x_1 + a_{12}x_2 + x_1g_{11}(x_1,x_2) + x_2g_{12}(x_1,x_2) \\ f_2(x_1,x_2) &= a_{21}x_1 + a_{22}x_2 + x_1g_{21}(x_1,x_2) + x_2g_{22}(x_1,x_2) \end{aligned}\right\} \qquad \text{(A.4,15)}$$

with the functions $g_{ij}(\ ,\ )$ continuous in a neighborhood of the origin and $g_{ij}(0,0) = 0$ $(i, j = 1, 2)$. Note that polynomial equations, such as given by (A.4,14) in Question A.4,12, can always be written in the form (A.4,15), and hence satisfy the more stringent conditions: their local phase portrait is just like that of their linearization.

Question A.4,16. The reader will verify in equation (A.4,15) that $D\vec{f}(\vec{0}) = A$ entails $g_{ij}(0,0) = 0$. Also continuity of $\vec{f}(\vec{x})$ entails continuity of $g_{ij}(x_1,x_2)$, except at $\vec{0}$. The continuity of $g_{ij}(\ )$ at $\vec{0}$ follows from the continuity of $D_j f_i(\ )$.

Question A.4,17. Putting

$$f_i(x_1,x_2) - a_{i1}x_1 - a_{i2}x_2 = \chi_i(x_1,x_2), \quad (i = 1, 2)$$

where $a_{ij} = D_j f_i(0,0)$, we find

$$\chi_i(x_1,x_2) = \int_o^1 \frac{d}{dt}\chi_i(tx_1,\ tx_2)\,dt$$

$$= \int_o^1 D_1\chi_i(tx_1,\ tx_2)\cdot x_1\,dt + \int_o^1 D_2\chi_1(tx_1,\ tx_2)\cdot x_2\,dt$$

$$= x_i g_{i1}(x_1,x_2) + x_2 g_{i2}(x_1,x_2),$$

where the g_{ij} denote certain obvious integrals. Trace exactly how the continuity of $g_{ij}(\ )$ follows from what smoothness properties of $\chi_i(\ )$, hence of $f_i(\ )$.

Question A.4,18. With $\vec{f}(\vec{x})$ as defined by (A.4,15), and $g_{ij}(0,0)$ equal to zero and continuous, show that $D_j f_i(0,0) = a_{ij}$.

Theorem A.4,4. Consider the differential equation ($\vec{x}$ being a 2-vector)

$$\dot{\vec{x}} = \vec{f}(\vec{x}),$$

with $\vec{f}(\vec{0}) = \vec{0}$. Let $D\vec{f}(\vec{0}) = A$ have two unequal, strictly negative eigenvalues, say $\underline{\lambda_2 < \lambda_1 < 0}$, but let $\vec{f}(\vec{x})$ satisfy equality (A.4,15). Then as $t \uparrow \infty$, every orbit of $\dot{\vec{x}} = \vec{f}(\vec{x})$ tends to the singular point $\vec{0}$ in one of four specific directions, an infinite number of them in the direction of $\vec{v}_1$ and $-\vec{v}_1$, and exactly one of them in the direction of $\underline{\vec{v}_2}$, and one in the direction of $\underline{-\vec{v}_2}$. Here $\vec{v}_1$ is the eigenvector of $D\vec{f}(\vec{0})$ corresponding to λ_1, and $\vec{v}_2$ is the eigenvector of $D\vec{f}(\vec{0})$ corresponding to λ_2.

Remark A.4,3. If in Theorem A.4,3 the two strictly negative eigenvalues are allowed to be equal, then Question A.4,4 shows that the singular point $\vec{0}$ may well be a spiral point for $\dot{\vec{x}} = \vec{f}(\vec{x})$. However, if the two strictly negative eigenvalues in Theorem A.4,4 are equal, then the singular point 0 will be a proper node for $\dot{\vec{x}} = \vec{f}(\vec{x})$ just as much as for $\dot{\vec{x}} = D\vec{f}(\vec{0})\cdot\vec{x}$. The difference between the two theorems is, of course, that Theorem A.4,4 imposes a stricter degree of smoothness on $\vec{f}(\vec{x})$. (The statements made in this Remark, though quoting conditions listed in previous theorems, don't follow from those theorems.)

The last theorem we will quote concerns saddle points:

Theorem A.4,5. Consider the differential equation ($\vec{x}$ being a 2-vector)

$$\dot{\vec{x}} = \vec{f}(\vec{x}),$$

with $\vec{f}(\vec{0}) = \vec{0}$. Let $D\vec{f}(\vec{0}) = A$ have two *unequal*, real *eigenvalues* with

$$\underline{\lambda_1 < 0 < \lambda_2},$$

and let $\vec{f}(\vec{x})$ satisfy equality (A.4,15). Then as $t \uparrow \infty$, *exactly one* orbit of $\dot{\vec{x}} = \vec{f}(\vec{x})$ tends to the singular point $\vec{0}$ in the direction of $\vec{v}_1$, and another in the direction of $-\vec{v}_1$. Again as $t \downarrow -\infty$, exactly one orbit of $\dot{\vec{x}} = \vec{f}(\vec{x})$ tends to the singular point $\vec{0}$ in the direction of $\vec{v}_2$, and another in the direction of $-\vec{v}_2$. All the other orbits eventually get away from $\vec{0}$, both as $t \uparrow \infty$ and as $t \downarrow -\infty$. In other words, a *saddle point* remains a saddle point when we go from a linear system to a nonlinear one under the above conditions, and the initial directions of the separatrices are determined just as in the linear systems.

Remark A.4,4. All the above theorems and discussions were concerned with the point $\vec{0}$ as a singular point of

$$\dot{\vec{x}} = \vec{f}(\vec{x}).$$

Now suppose $\vec{\alpha}$ is a singular point of the same equation, i.e., $\vec{f}(\vec{\alpha}) = \vec{0}$. Then put $\vec{x} = \vec{y} + \vec{\alpha}$: it follows that

$$\dot{\vec{y}} = \vec{f}(\vec{\alpha} + \vec{y}) = \vec{g}(\vec{y}), \quad \text{say.}$$

Consequently $\vec{g}(\vec{0}) = \vec{f}(\vec{\alpha})$, and $D\vec{g}(\vec{0}) = D\vec{f}(\vec{\alpha})$. Thus the linearization method re the singular point $\vec{\alpha}$ of $\dot{\vec{x}} = \vec{f}(\vec{x})$ amounts to the study of the matrix $D\vec{f}(\vec{\alpha})$ instead of $D\vec{f}(\vec{0})$, and of the 'nonlinear part' $\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot(\vec{x}-\vec{\alpha}) = \vec{g}(\vec{y}) - D\vec{g}(\vec{0})\cdot\vec{y}$ instead of $\vec{f}(\vec{x}) - D\vec{f}(\vec{0})\cdot\vec{x}$. Except for this modification all theorems remain valid for the study of singular points $\vec{\alpha}$.

Remark A.4,5. The reader should go over all five theorems and practice time-reversal, $t = -T$. This causes all eigenvalues to switch signs. Watch how this affects the conclusions.

Example. Consider the equation

$$\left\{\begin{aligned} \dot{x}_1 &= x_1(\epsilon_1 - \alpha_1 x_1 - \beta_1 x_2) \\ \dot{x}_2 &= x_2(\epsilon_2 + \alpha_2 x_1 - \beta_2 x_2) \end{aligned}\right\}, \quad \left\{\begin{aligned} &0 < \epsilon_2\beta_1 < \epsilon_1\beta_2 \\ &0 < \alpha_i \ ; \ 0 < \beta_i \end{aligned}\right\}.$$

Singular points: $(0,0)$, $(0, \epsilon_2/\beta_2)$, $(\epsilon_1/\alpha_1, 0)$, (ξ_1, ξ_2), where

$$\left\{\begin{aligned} \epsilon_1 - \alpha_1\xi_1 - \beta_1\xi_2 &= 0 \\ \epsilon_2 + \alpha_2\xi_1 - \beta_2\xi_2 &= 0 \end{aligned}\right\} \Rightarrow \xi_1 > 0 \ \text{ and } \ \xi_2 > 0.$$

The reader will find $D\vec{f}(\vec{\alpha})$ at each of these four singular points, check the nature of the relevant nonlinear parts, and determine the asymptotic directions or separatrices, as the case may be. Partial answer: in the order given the above four singular points are an unstable node, a saddle point, a saddle point, and a stable node.

APPENDIX B

BIBLIOGRAPHY

Section 1.3 presents a short list of journals in which ongoing research is published that should be of interest to the audience. Section 1.3 also gives a short list of fairly recent books in the general area of ecosystems. With respect to the other titles below, most of them are quoted simply to recognize the source of certain statements or ideas discussed in the text. Those are not at all crucial to the successful teaching of this module, and the instructor need not hunt for them unless he or his class are wanting to expand upon the topic for which they are quoted. At the end of each title we listed the section(s) in which it is mentioned. A very few books are essential or very desirable to have around as a supplement to the written text. Most of these are available in fairly cheap paperbacks; they are indicated by the words *essential* or *very desirable*.

Aggarwal, J. K. (1972), Notes on nonlinear systems; New York, Van Nostrand Reinhold (in the paperback series 'Notes on System Sciences'); 214 pp. One of the titles mentioned at the end of Remark 2.1,5 (in section 2.1); it would be very desirable to have one of these around, and the present one is an inexpensive paperback.

Allee, W. C.; Emerson, A. E.; Park, O.; Park, T.; Schmidt, K. P. (1949). Principles of animal ecology; Philadelphia, W. B. Saunders; xii + 837 pp. See sections 1.1 and 3. If the library on your campus has it, it is worthwhile for you and your students to have a good look at it, but don't have it ordered.

d'Ancona, U. (1954), The struggle for existence (trans. by A. Charles and R. F. J. Withers); Leiden, E. J. Brill; vol. 6 of Bibliotheca Biotheoretica xi + 274 pp. Available in hardcover and in paperback. See sections 1.1, 1.2, 2.1 (Question 2.1,3) and 4. It is *very desirable*, in fact almost *essential* to have this book available (for the instructor).

Andronov, A. A.; Leontovich, E. A.; Gordon, I. I.; Maier, A. G. (1973). Qualitative theory of second-order dynamic systems; New York, Wiley (a Halsted Press book); xxiii + 524 pp. (trans. from a 1966 Russian edition). One of the titles mentioned at the end of Remark 2.1,5 (in section 2.1); it would be very desirable to have one of these around; the present one is unusually detailed and exhaustive, but very expensive, and it does not discuss systems of dimension larger than two.

Ayala, F. J. (1970), Competition, coexistence, and evolution; pp. 121-158 in: M. K. Hecht and W. C. Steere (eds.) (1970), Essays in evolution and genetics in honor of Th. Dobzhansky: a supplement to evolutionary biology; Apple-Century-Crofts/Meredith; xv + 594 pp. See section 3.3.

Balogh, J. (1958), Lebensgemeinschaften der Landtiere; ihre Erforschung unter besonderer Berücksichtigung der zoozönologischen Arbeitsmethoden; Budapest, Verlag der Ungarischen Akademie der Wissenschaften, and Berlin, Akademie-Verlag; 560 pp. See section 1.1.

Bartlett, M. S. (1961), Monte Carlo studies in ecology and epidemiology; pp. 39-55 in: Proc. Fourth Berkeley Sumposium on Mathematical Statistics and Probability, vol. 4 (Contributions to biology and problems of medicine), ed. by J. Neyman; University of California Press. See section 1.5.

Boudart, M. (1968), Kinetics of chemical processes; Englewood Cliffs, N. J., Prentice-Hall; ix + 246 pp. See section 2.1.

Brauer, F. and Nohel, J. A. (1969), The qualitative theory of ordinary differential equations; an introduction; New York, Benjamin (now Reading, Addison-Wesley); xi + 314 pp. One of the titles mentioned at the end of Remark 2.1,5 (in section 2.1); it would be very desirable to have one of these around; the present one is hardcover, but not exceedingly expensive.

Coppel, W. A. (1966), A survey of quadratic systems; Journal of Differential Equations 2, pp. 293-304. See section 4 (Question 4.8).

Ellenberg, H. (ed.) (1971), Integrated experimental ecology; methods and results of ecosystem research in the German Solling project; New York, Springer; xx+ 214 pp. (of which pp. 1-15 carry an introductory survey by H. Ellenberg). See section 1.1. The book is vol. 2 of the series, 'Ecological Studies, Analysis and Synthesis."

Ėman, T. I. (1966), O nekotoryh matematiceskih modeliah biogeocenozov; Problemy Kibernetiki (pod red. A. A. Liapunova) 16, str. 191-202; Moskva, Izd. Nauka. See sections 2.2 and 4 (Question 4.7).

Feller, W. (1939), Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung; Acta Biotheoretica (Leiden) 5, pp. 11-40. See section 1.5. A famous paper, often quoted as starting the 'stochastization' of deterministic models, but not essential.

Feynman's Lectures in Physics (1963, repr. 1965), vol. 1; Reading, Mass., Addison-Wesley; 52 chapters. See section 1.5, point γ. The sections 8-2, 8-3, and 9-5 of Feynman's book are a very good specimen of the classical argument for the derivative as a measure of speed or rate.

Flanders, S. E. and Badgley, M. E. (1963), Prey-predator interactions in self-balanced laboratory populations; Hilgardia 35, pp. 145-183. See section 2.1.

Gause, G. F. (1934, 1971), The struggle for existence; Williams and Wilkins; paperback edition by Dover in 1971; ix + 163 pp. Very desirable. This paperback is well worth the price. The book gives a number of graphs representing data. See sections 1.2, 1.5, 2.2, 3.3.

Gause, G. F. (1935), Experimental demonstration of Volterra's periodic oscillations in the number of animals; The Journal of Experimental Biology 12, pp. 44-48. See section 1.2. This article has graphs of the same type as the above book, but with experimentally controlled 'reproduction' rates.

Gilpin, M. E. and Justice, K. E. (1972), Reinterpretation of the invalidation of the principle of competitive exclusion; Nature 236, pp. 273-301. See section 3.3, also Remark 3.3,1.

Goel, N. S.; Maitra, S. C.; Montroll, E. W. (1971), On the Volterra and other nonlinear models of interacting populations; New York, Academic Press; vii + 145 pp.; Reprint from Reviews of Modern Physics 43 (1971), pp. 231-276. See sections 1.3, 2.1 (Remark 2.1,6) and 4. Too advanced; also contains errors.

Gomatam, J. (1974a), A new model for interacting populations-I: two-species systems; Bulletin of Mathematical Biology 36, pp. 347-353. See section 2.2, euations (2.2,4) and (2.2,5).

Gomatam, J. (1974b), A new model for interacting populations-II: principle of competitive exclusion; Bulletin of Mathematical Biology 36, pp. 355-364. See sections 3.2 (Question 3.2,1) and 3.3.

Hardin, G. (1960), The competitive exclusion principle; Science 131, pp. 1292-1297. See section 3.3.

Hirsch, M. W. and Smale, S. (1974), Differential equations, dynamical systems, and linear algebra; Academic Press, xi + 358 pp. See section 2.1 (Discussion question 2.1,9).

Holling, C. S. (1966), The functional response of invertebrate predators to prey density; Memoirs of the Entomological Society of Canada 48, pp. 1-86. See section 2.2 (after Question 2.2,10). This paper could, together with Watt, 1959 (see below), form a basis of a student project concerning some details of the model building process.

Holling, C. S. and Ewing, S. (1971), Blind man's buff: exploring the response space generated by realistic ecological simulation models; pp. 207-233 in: G. P. Patil, E. C. Pielou and W. E. Waters (eds.), Statistical Ecology, vol. 2. The Pennsylvania State University Press. See section 2.2 (Discussion question 2.2,5).

Huffaker, C. B. (1958), Experimental studies on predation: dispersion factors and predator-prey oscillations; Hilgardia 27, pp. 343-383, reprinted as pp. 164-204 in: W. E. Hazen (ed.) (1964), Readings in population and community ecology; Philadelphia, Saunders x + 338 pp., available as paperback. See sections 1.5 (first paragraph), 2.2 (first paragraph), and 2.3 (Remark 2.3,2). Huffaker describes extensive experiments, highlighting certain tensions between reality and certain models discussed in our module. Hazen's book of "readings" has more papers of interest; it would be desirable to have this paperback available. Huffaker's paper is also quoted by Leigh, 1968, see below.

Hurewicz, W. (1964), Lectures on ordinary differential equations; Cambridge, Massachusetts, M. I. T. Press, xvii + 122 pp.; paperback. One of the titles mentioned at the end of Remark 2.1,5 (in section 2.1); it would be very desirable to have one of these around, and the present one is an inexpensive paperback (beautiful, but a bit more abstract than Aggarwal).

Iglehart, D. L. (1964), Multivariate competition processes; Ann. Math. Stat. 35, pp. 350-361; also see Iglehart, D. L. (1964), Reversible competition processes; Z. f. Wahrscheinlichkeitstheorie und verwandte Gebiete 2, pp. 314-331. See section 1.5 (second paragraph). The work on 'competition processes' is not really in the stage of biological applicability yet.

Jacobs, G. J. (ed.) (1966), Proceedings of a conference on theoretical biology; a conference, organized by the Amer. Inst. of Biol. Sci. under the sponsorship of NASA, at Princeton, N. J., November 1963; published by the Scientific and Technical Information Division of NASA under the number NASA SP-104; vii + 188 pp. See end of section 2.1.

Kakutani, S. and Markus, L. (1958), On the non-linear difference-differential equation $y'(t) = [A - B\,y(t-\tau)]\,y(t)$; pp. 1-18 in: S. Lefschetz (ed.) (1958), Contributions to the theory of nonlinear oscillations, vol. 4; Ann. of Math. Studies 41; Princeton University Press. See Discussion question 1.5,10 (in section 1.5).

Kerner, E. H. (1972), Gibbs ensemble: biological ensemble; the application of statistical mechanics to ecological, neural, and biological networks; New York, Gordon and Breach; xi + 167 pp. See section 2.1 (Remark 2.1,6). This book consists of four reprinted articles with a lengthy (66 pp.) and far from flawless introduction; the book will be too advanced for most of the audience.

Kolmogorov, A. N. (1972), Kacestvennoe izucenie matematiceskih modelei dinamiki populiacii (Qualitative investigation of mathematical models of population dynamics); pp. 101-106 in: Problemy Kibernetiki 25; Moskva, Nauka. See section 2.1, starting after Question 2.2,2 and continuing for three or four pages. All that is needed for the questions on those pages is spelled out, so the original reference is not needed for teaching this material.

Lapidus, L. and Seinfeld, J. H. (1971), Numerical solution of ordinary differential equations; Academic Press, xii + 299 pp. See section 1.6 (part f of Question 1.6,3).

Leigh, E. G., Jr. (1968), The ecological role of Volterra's equations; pp. 1-61 in: Lectures on Mathematics in the Life Sciences (the first volume in the series); Proceedings of the Symposium on Mathematical Biology held at Washington, D. C., December, 1966; Providence, R. I., American Mathematical Society; vi + 117 pp. See section 1.4, last paragraph. The reason for quoting this paper consists in the tables given on pages 58 through 60 of that paper. However, upon comparing the numbers in these tables with the graphs from which E. G. Leigh took them, it becomes apparent that the accuracy of the process was low. So it is better that the reader himself look at these graphs and obtain these numbers if he wants them. The graphs are to be found in Huffaker, 1958 (see above), Figure 18, and in MacLulich, 1937 (see below), Figures 16 and 17 (pp. 109 and 110).

Leigh, E. G., Jr. (1971), Adaptation and diversity; natural history and the mathematics of evolution; San Francisco, Freeman, Cooper and Company; 288 pp. See section 2.3 (Remark 2.3,2). Its Figure 7-5 is the same as Rosenzweig's Figure 1 (see below). What the figure does, should rather be done by the students themselves: take a graph from Huffaker (1958), and from these data construct the size of the horizontal and the vertical arrows and their components. Thus one obtains a phase portrait.

Leslie, P. H. (1958), A stochastic model for studying the properties of certain biological systems by numerical methods; Biometrika 45, pp. 16-31. See sections 1.6 (Discussion question 1.6,6) and 2.2 (in the paragraph preceding Question 2.2,7, and in Discussion question 2.2,9). See remark re Leslie and Gower (1960) below.

Leslie, P. H. (1962), A stochastic model for two competing species of Tribolium and its application to some experimental data; Biometrika 49, pp. 1-25. See section 1.5 (second paragraph). See remark re Leslie and Gower (1960) below.

Leslie, P. H. and Gower, J. C. (1960), The properties of a stochastic model for the predator-prey type of interaction between two species; Biometrika 47, pp. 219-234. See section 1.5 (second paragraph). The above two papers by Leslie and the present one are typical for a number of publications in that time from the Bureau of Animal Population in Oxford.

Lotka, A. J. (1907), Studies on the mode of growth of material aggregates; American Journal of Science, 4th. Ser., 24 (whole number 174), pp. 199-216, with an addendum on pp. 375-376. See section 1.2 (second paragraph).

Lotka, A. J. (1925, 1956), Elements of physical biology; Baltimore, Williams and Wilkins (1925); reprinted as a paperback, Elements of mathematical biology, Dover (1956), xxx + 465 pp. See sections 1.2 (second paragraph), and 2.1 (first paragraph). Somewhat desirable, available as an inexpensive paperback.

MacArthur, R. H. and Connell, J. H. (1966), The biology of populations; Wiley, xv + 200 pp. See section 2.2 (first paragraph). Of this book Chapter 6 is closest to our present concerns.

MacArthur, R. H. (1972), Geographical ecology; patterns in the distribution of species; New York, Harper and Row, 269 pp. See section 1.3.

MacLulich, D. A. (1937), Fluctuations in the numbers of the varying hare (Lepus americanus); Toronto, University of Toronto Press (No. 43 of The University of Toronto Studies, Biological Series); 136 pp. See remarks made regarding Leigh (1968) above.

Mamdani, M. (1972), The myth of population control; family, caste, and class in an Indian village; New York and London, Monthly Review Press; 173 pp. See Appendix, remarks made regarding Discussion question 1.5,1. This book gives a detailed discussion of the cultural and economic barriers against birth control in a typical Indian village.

May, R. M. (1973a), Stability and complexity in model ecosystems; Princeton University Press; ix + 235 pp. See sections 1.3, 2.2 (paragraph preceding Discussion question 2.2,11), and 3.1. Section 1.3 gives a short description of the book. It is very desirable for the instructor to have it around.

May, R. M. (1973b), On the relationships among various types of population models; The American Naturalist 107, pp. 46-57. See section 1.6 (end of first paragraph).

Maynard Smith, J. (1968), Mathematical ideas in biology; Cambridge University Press; vii + 152 pp. See section 1.3, where a short characterization is given.

Maynard Smith, J. (1974), Models in ecology; Cambridge University Press, xii + 146 pp. See section 1.3, where a short characterization is given.

Mertz, D. B. (1972), The Tribolium model and the mathematics of populations growth; pp. 51-78 in: Annual Review of Ecology and Systematics, vol. 3; 102 references. See sections 1.3, 1.5 (end of second paragraph), and 3.2 (Remark 3.2,1). If this volume is in the campus library, it is quite worth while to take a good look at this article.

Neyman, J.; Park, T.; Scott, E. L. (1956), Struggle for existence; the Tribolium model; biological and statistical aspects; pp. 41-79 in: Proc. of the Third Berkely Symposium on Mathematical Statistics and Probability (held December 1954 and July-August 1955); vol. 4, Contributions to Biology and Problems of Health; ix + 179 pp. See section 1.5 (end of second paragraph) and 3.2 (Remark 3.2,1). Emphasis is on how the stochastic model is different from the deterministic model, and how it is similar to it.

Pearl, R. and Reed, L. J. (1920), On the rate of growth of the population of the United States since 1790 and its mathematical representation; Proc. of the National Academy of Sciences 6, pp. 275-288. See section 1.5 (Discussion question 1.5,9).

Poletaev, I. A. (1966), O mathmaticeskih modeliah elementarnyh processov v biogeocenozah; Problemy Kibernetiki (pod red. A. A. Liapunova) 16, str. 171-190; Moskva, Izd. Nauka. See sections 2.2 (paragraph after Discussion question 2.2,9) and 4 (Question 4,7).

Problemy Kibernetiki, an ongoing serial publication, edited by A. A. Liapunov, published in Moscow by 'Nauka,' on any subject related to cybernetical problems. Volume 25 (1972), is especially heavy on population dynamics. It may be available in translation. See section 2.2 (paragraph preceding Discussion question 2.2,11).

Reuther, G. E. H. (1961), Competition processes; pp. 421-430 in: Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, (Probability Theory); University of California Press. See section 1.5 (second paragraph). The work on 'competition processes' is not really in the stage of biological applicability yet.

Rosenzweig, M. L. (1969), Why the prey curve has a hump; The American Naturalist 103, pp. 81-87.

Royama, T. (1971), A comparative study of models for predation and parasitism; Researches on Population Ecology, Supplement no. 1, September 1971; The Society of Population Ecology, c/0 Entomological Laboratory, Kyoto University, Kyoto, Japan; 91 pp. See section 2.2 (paragraph preceding Discussion question 2.2,11). Theoretically oriented; would be useful for a far more detailed course.

Stewart, F. M. and Levin, B. R. (1973), Partitioning of resources and the outcome of interspecific competition: a model and some general considerations; The American Naturalist 107, pp. 171-198. See section 3.3 (second paragraph).

Tansley, A. (1935), The use and abuse of vegetational concepts and terms; Ecology 16, pp. 284-307. See section 1.1.

Tsokos, C. P. and Hinkley, S. W. (1973), A stochastic bivariate ecology model for competing species; Mathematical Biosciences 16, pp. 191-208. See section 1.5 (second paragraph).

Usher, M. B. and Williamson, M. H. (eds.) (1974), Ecological stability; London, Chapman and Hall, and New York, Wiley; 196 pp. See section 1.3.

Vaart, H. R. van der (1973), A comparative investigation of certain difference equations and related differential equations: implications for model-building; Bulletin of Mathematical Biology 35, pp. 195-211. See section 1.6. Essential if section 1.6 is used for a student project.

Volterra, V. (1927), Variazioni e fluttuazioni del numero d'individui in specie animali conviventi; Atti della R. Accademia Nazionale dei Lincei, Anno 324; Serie sesta, Memorie dellla classe di scienze fisiche, matematiche e naturali, volume 2, pp. 31-113. See sections 1.2 and 2.1 (Remark 2.1,1). This title is almost the same as the title of the article published in the Rendiconti Comitato Talassografico Italiano (see section 1.2). There are no observational data in the present title; only mathematical theory.

Volterra, V. (1929, 1959), Theory of functionals and of integral and integro-differential equations; Blackie and Sons (1929), Dover paperback (1959); [39] + xiv + 226 pp. See section 1.2 (first paragraph). The separately numbered pages, [5] through [39], contain a good biography, written by E. Whittaker, especially pp. [19]-[25], [32]-[33] and [36]-[39].

Watt, K. E. F. (1959), A mathematical model for the effect of densities of attacked and attacking species on the number attacked; The Canadian Entomologist 91, pp. 129-144. See section 2.2 (paragraph preceding Discussion question 2.2,11. See remark made above regarding Holling (1966).

Watt, K. E. F. (1968), Ecology and resource management; a quantitative approach; McGraw-Hill; xii + 450 pp. See section 2.2 (paragraph preceding Discussion question 2.2,11).

Williamson, M. (1972), The analysis of biological populations; London, Arnold and New York, Crane and Russak; 180 pp. See section 1.3, where a short characterization is given.

APPENDIX C

HINTS, REMARKS RE QUESTIONS

Some of the questions, especially the discussion questions, could puzzle the instructor or class as to what kind of answer should be expected. It is hoped that the following remarks about some of them may alleviate such puzzlements. These should be especially helpful to persons with scant experience in applications of mathematics to biology. The following remarks go by the number of the questions.

Discussion question 1.5,1. In human populations, for instance, behavioral and cultural patterns that were established during an era of relatively low population density and primitive medical knowledge have a tendency to change slower than might be desirable for the standpoint of increased population density. So family size, or equivalently percentage-wise rate of increase, is not really determined by today's population density, but by that of a century ago, say. For example, see the difficulties encountered by "family planning" or "population control" in India, as discussed by Mamdani (1972). On a different time scale, long gestation periods introduce delays in the conversion of food into increases in population numbers. Students or instructors, especially those with a biological background, might well come up with more examples.

Discussion question 1.5,2. Answers: $\varphi(x,y)$ and $\psi(x,y)$, any pair of functions independent of h.

Question 1.5,4. Any of the following methods will be fine; it would be useful to show more than one. (1) Substitute (1.5,5d) into (1.5,5c); this also decides the question of existence of a solution. (2) Divide both members of (1.5,5c) by x^2 and change variables according to

$$y(t) = \frac{1}{x(t)};$$

this results in a linear equation in $y(t)$. (3) Multiply both members of

(1.5,5c) by $K/[x(K-x)]$; the use of the method of partial fractions will then yield the differential equation

$$\frac{\dot{x}(\square)}{x(\square)} + \frac{\dot{x}(\square)}{K - x(\square)} = \gamma K;$$

now integrate with respect to $\square$ from t_0 to t. The third method is the most delicate one:

$$\int_{t_0}^{t} \frac{x(\square)}{x(\square)}\, d\square = \log \left| \frac{x(t)}{x_0} \right| ,$$

and one has to worry about $x(t)$ not taking the value zero, so that $x(t)/x_0$ will have a constant sign and the absolute value bars can be deleted.

Discussion question 1.5,5. Experience shows that some audiences have a quick understanding of this question, others wrestle with it. At a more intuitive level, the point is that it is not true that every quadruple of values (t_0, x_0, γ, K) with $0 < \gamma$, $0 < K$, $0 < x_0 < K$, determines a different function of the form (1.5,5d). At a formal level the question can be rephrased as follows: show there is not a 1-1-correspondence between the points of 4-dimensional (t_0, x_0, γ, k)-space (with $0 < \gamma$, $0 < K$, $0 < x_0 < K$) and the collection of all logistic curves of the form (1.5,5d); rather there is a 1-1-correspondence between the points of 3-dimensional (A, B, C)-space (with $0 < A$, $0 < B$, $0 < C$) and the collection of all logistic curves, as can be seen by rewriting (1.5,5d) as

$$x(t) = \frac{A}{1 + B\, e^{-Ct}}$$

Specifically, the individual values of x_0 and t_0 do not matter: the only thing that counts is the value of

$$\frac{K - x_0}{x_0}\, e^{+\gamma K t_0}$$

The question is important in view of statistical estimation of parameters: (t_0, x_0) is not identifiable.

Question 1.5,6. It is recommended that the students draw more than one graph: two or three in the strip $0 < x < K$, one or two in the halfplanes $x > K$ and $x < 0$.

Discussion question 1.5,9. Re the question at the very end. Verhulst said that his 'agricultural' argument does lead to an expression for $f(x)$ of the form $K-x$, but it does not really. According to the verbal argument given, the population grows purely exponentially until the level $x = b$ is reached (at time t_b, say). For $x > b$ the differential equation model is of the form

$$\frac{\dot{x}}{x} = (r - m\cdot(x-b)) = r + mb - mx$$
$$= m(K - x),$$

with $mK = r + mb$, and $x(t_b) = b$. The curve that answers Verhulst's description consists, therefore, of two parts, each part described by a different formula.

Discussion question 1.5,10. The rationale for the equation displayed here is as follows. Think of a human population in which things like family size are decided on the basis of traditional attitudes (cf. remarks re Discussion question 1.5,1 above). Then $\tau \sim 100$ years, say: the factor in the formula that is supposed to represent increased procreational caution, viz. $(K-x)$, should be evaluated 100 years before t; this is the factor that should reflect attitudes about family size, and these attitudes were formed under the influence of the population size of τ years ago. On the other hand, the number of families now is more or less proportional to the population size now, or, at least, to the population size of a much shorter time ago than the above-postulated period of adjustment of attitudes to population size. The same equation is proposed with essentially the same argument (though less elaborate) by Kakutani and Markus (1958). Of course, in a population in which family size would be determined less by tradition than by the experiences and feelings of the parents during their childhood, and in which the bulk of people are having children during the same relatively short age span, a more logical model would be

$$x(t) = \gamma\cdot x(t-\tau)\cdot[K - x(t-\tau)] \qquad \text{(a)}$$

and

$$x(t) = \gamma\cdot x(t-\tau)\cdot[K - x(t-\tau)], \qquad \text{(b)}$$

it will become clear (as soon as you try to do it) that the knowledge of $x(t_o)$ is now not enough to get started: $\dot{x}(t_o)$ cannot be computed from $x(t_o)$, not for equation (a) and not for equation (b): one needs to know the function $x(\ )$

over the time interval from $t_o - \tau$ to t_o: an arc instead of a point. Suppose this initial arc represents a monotone increasing function of t, then the solution of equation (a) will, at any level x up to K, be steeper than any logistic with the same γ and K. Suppose this initial arc represents a piece of a regular logistic with the same γ and K, then the solution of equation (b) will be obtained by translating a suitable logistic (with that γ and K) upwards over the distance $x(t_o) - x(t_o - \tau)$. Finally, going back to equation (a), suppose that $x(t') = K$; then $x(t' - \tau) < K$ and $\dot{x}(t') > 0$. So the population size will increase beyond K, and $\dot{x}(\ )$ will be zero only at $t' + \tau$. Thereafter $\dot{x}(\ ) < 0$ for a while. Pursuing this argument, one can make the oscillations of $x(t)$ around K as shown by Kakutani and Markus, 1968, very plausible indeed.

Questions in section 1.6. If you want to let the students do some work on this subject matter, write for a reprint of H. R. van der Vaart (1973).

Question 2.1,1. Choose β_1 and β_2 such that $p\,\beta_2 = q\,\beta_1$.

Question 2.1,2. See sections A.1 and A.2 in Appendix A.

Question 2.1,4. The intention of this question should be clear by comparison with Question 2.1,1. The point is: no scale change of the form $x = \beta_1 x_1$, $y = \beta_2 x_2$ will be able to turn the matrix

$$\begin{Bmatrix} \delta & -p \\ +q & -\zeta \end{Bmatrix}$$

into a skew-symmetric one. The reason for this is: a skew-symmetric matrix has diagonal elements equal to zero, and neither $-\delta\,\beta_1$, nor $-\zeta\,\beta_2$ will ever be zero, since δ and ζ are given as positive numbers, and $\beta_i = 0$ would make the change meaningless.

Question 2.1,5. See sections A.1 and A.2.

Question 2.1,6. See section A.4 of Appendix A, especially the final example in that section. One should not find any contractions. But some of the things conjectured from the arrow method should now become more precise and certain.

Question 2.1,7. There are too many non-zero coefficients in equation 2.1,5. So we cannot now get rid of all the unwanted terms and thus separate variables.

Discussion question 2.1,8. Looking at (2.1,5) we can say that the ratio δ/γ_1 indicates how much worse the presence of one prey individual makes the environment for the prey species, whereas the ratio q/γ_2 indicates how much better the presence of one prey individual makes the environment for the predator. The model yields coexistence of predator and prey iff the percentage deterioration for the prey is less than the percentage improvement for the predator. In the opposite case the advantage for the predator to have the prey around is apparently not enough to keep the predator in business. Now this sounds as if it would make the biologist moderately happy. The only thing that the biologist could object to is: why does ζ not play a role in the decision re coexistence of the two species? It does to some extent: it helps determine the size of the predator population at equilibrium; but other than that I don't see a ready answer to this remark, and it is a bit odd for the crowding effect in the predator not to be terribly important in the question of coexistence, while the crowding effect in the prey is that important. Another possible question would be: why should q be considered relative to γ_2? Well, γ_2 describes how much the predator needs just to stay in business.

Discussion question 2.1,9. The locus of all points with $\dot{x} = 0$ should be a horizontal line through the singular point. The locus of all points with $\dot{y} = 0$ should be ..?.. (see equation (2.1,2)). And neither is the case in this diagram.

Question 2.2,1. Any trajectory that crosses that portion of the isocline, reaches its left-most point right there, at that crossing point. Thereafter, the danger for the prey to decrease more has passed. So if that isocline keeps far enough away from the y-axis, the prey will not become extinct.

Question 2.2,2. Somewhat similar to preceeding question: the prey can't $\uparrow +\infty$.

Question 2.2,3. The one term which makes equation (2.1,5) different from equation (2.2,1) is obviously $y\cdot(\ldots - \zeta y)$. It contradicts Kolmogorov's assumption that the predators act independently of each other (see point (1) in the latter half of Question 2.2,3).

Discussion question 2.2,4. If the equation $K_2(x) = 0$ has any roots, then the 'horizontal' isocline is the union of the x-axis $(y = 0)$ and any vertical line $x = \xi$, where ξ is a root of $K_2(x) = 0$. Also any singular point must belong to the 'horizontal' isocline. All of Kolmogorov's sketches portray only one singular point outside the coordinate axes. So the only spot on any trajectory (outside the x-axis) where that trajectory can have a horizontal tangent line, must be on the vertical line through that lonely singular point. Most of Kolmogorov's sketches lack this property.

Discussion question 2.2,5. The y-axis belongs to the 'vertical' isocline iff $L(0) = 0$.

Discussion question 2.2,6. $K_1(x)$ is the relative rate of increase of the prey in the absence of the predator. The assumption that it is monotone decreasing represents the crowding effect and is therefore not unreasonable (except perhaps for extremely small values of x); whether $K_1(x)$ should be negative for large enough x is something else. The assumptions re $K_2(x)$ and $L(x)$ are in line with our earlier discussions, except maybe for $L(0) > 0$; this inequality means that a prey population of size zero keeps decreasing as long as there are some predators left, which sounds odd (whoever saw a population of size less than zero?). The equation of the 'vertical' isocline is

$$y = \frac{K_1(x)}{L(x)} \cdot x.$$

From the definition of $K_1(\ )$ and $L(\ )$ it is clear that these functions have some positive value for each value of x less than the root of $K_1(x) = 0$. So for each value of x less than that root there is a value of y (viz., $x\cdot K_1(x)/L(x)$) such that the trajectory through (x,y) has a vertical tangent line at (x,y). Looking at the diagrams one sees that several of them do not show this property (and are, hence, erroneous). There are several other difficulties with these diagrams. The author of this module does not see, for instance, how the singular points can possibly be located as in diagram number 2:

the singular point $(A,0)$ must be on the 'vertical' isocline, so $K_1(A)\cdot A - L(A)\cdot 0 = K_1(A)\cdot A = 0$; therefore $K_1(A) = 0$, so $K_1(x) < 0$ for $x > A$. The singular point (B,C) must also be on the 'vertical' isocline, so $K_1(B)\cdot B - L(B)\cdot C = 0$. However, for $B > A$ and $C > 0$ it is clear that $K_1(B)\cdot B - L(B)\cdot C < 0$. These remarks should give a good idea of what all can be uncovered with respect to these diagrams, and we will not spoil your fun by spelling it all out. Note that for some of the questions you need to use the linearization method as discussed in section A.4 of Appendix A.

Question 2.2,7. The singular points are the origin (unstable node) and a point with both coordinates positive (saddle point). Let $\vec{p} = (p_1,p_2)$ be the second singular point. Then it is easy to express p_1 and p_2 in terms of the a, the b, and the r. but you do not need to do this in order to show that $\vec{p}$ is a saddle point. Note that $r_1 - a_1p_1 - b_1p_2 = 0$ and show the equation for the eigenvalues of the matrix $D\vec{f}(\vec{p})$ (cf. section A.4) is

$$\lambda^2 - (r_2 - a_1p)\lambda - p(a_1r_2 + a_2b_1) = 0.$$

So the product of the two eigenvalues is negative, which is possible only if they are real and have opposite signs (if λ_1 and λ_2 are complex, then they are conjugate complex and their product is positive). You can also find the directions of the separatrices of the saddle point at the saddle point. Note that for this equation the x-axis is not part of the 'horizontal' or the 'vertical' isocline. Also note that all trajectories that start above one of the separatrices (which one?) depart for $y \uparrow +\infty$, which is biologically unacceptable. So this model is not much of an improvement.

Question 2.3,1. The four answers, in the order of their corresponding questions are: 0, 0, positive, negative.

Discussion question 2.3,2. At any point on the 'vertical' isocline $\psi(x,y)$ (cf. equation (2.3,1)) is non-zero (unless that point is also on the 'horizontal' isocline), that is, $\psi(x,y)$ is either positive or negative. Because of the continuity of $\psi(x,y)$ it will have the same sign at points off the isocline as at points on the isocline, provided they are close enough. "Close enough" means here: "at the same side" of the 'horizontal' isocline, because $\psi(x,y)$ will not be able to change sign without hitting zero; and $\psi(x,y)$ can hit zero only on the 'horizontal' isocline (why?). Finally, when $(x,y) > 0$, then $\dot{y} > 0$ and the vertical arrow points upwards; when $\psi(x,y) < 0$, then $\dot{y} < 0$ and the vertical arrow points downwards. A similar story applies to horizontal arrows. This should enable the reader to explain all the arrows in Figure 2.3,4.

Discussion question 2.3,3. The two possibilities depicted in the two diagrams on the left seem to allow populations to explode, all the way to infinity. This would be biologically unacceptable: even insect pests eventually come down.

Question 2.3,4. At least one saddle point, and one stable or unstable spiral point.

Question 3.2,1. The equation $\varepsilon - \alpha_1 \log x - \alpha_2 \log y = 0$ determines y as a convex function of x (e.g., show that $y'' > 0$!). This opens the possibility for the coexistence point to be in the interior of the triangle formed by the origin, the singular point on the x-axis, and the singular point on the y-axis.

Chapter 7
POPULATION MATHEMATICS

Frank C. Hoppensteadt
New York University

1. Introduction

The estimation of the size and age composition of various biological populations is a significant concern to many decision makers. For example, among those who have a need for such information are the planners for the utilization of renewable natural resources such as fish and waterfowl, and the long range educational and economic planners concerned with demographic data of human populations. Although the ideas and techniques introduced in this unit have obvious utility for a whole set of biological populations, we will use the terminology of human population growth just to have a specific example at hand.

Attempts to use mathematics to describe the growth of human populations go back at least as far as the late 18th century, and one of the most familiar elementary models was proposed by the English economist and demographer T. R. Malthus (1766-1834). The Malthus model is based on the assumption that the instantaneous rate of growth of a population is proportional to the size of the population. Here and in the future the size of a population is assumed measured in convenient continuously divisible units (biomass or a similar quantity) so there is no inherent difficulty in using differential equations. Malthus' model leads to the prediction of unlimited population sizes; a conclusion which leads one to view the underlying assumptions with some skepticism. Various modifications of the Malthus model have been proposed which remove the objectionable conclusion. For example, a model proposed by a 19th century Belgian sociologist P. F. Verhulst and subsequently rediscovered by others leads to predicted population sizes which tend asymptotically to a constant value as time increases. However, the Verhulst model and many more recent ones, some of which lead to differential-delay equations or functional-differential equations of considerable mathematical interest, do not consider directly the influence of the age composition of a population on its development. This will be the primary concern of this module.

We have a number of decisions to make. First of all, shall we view the situation as a deterministic or a stochastic one? Obviously the biological-social situation we are modeling induces random events. However, if we are concerned with large populations, then we have some confidence that a deterministic model will provide a good first approximation. Also, the mathematical techniques required for an analysis of the deterministic model proposed here are less sophisticated than those which would be needed for the analysis of a stochastic model of similar generality. The choice of model is ultimately a decision of the investigator, and we choose a deterministic model.

Our second decision is whether or not to consider the spatial as well as the temporal variation of the population. For various reasons, among these a lack of adequate data for spatially distributed populations, we shall restrict our attention to the changes in the size and age composition of the population through time.

Our last decision is whether time should be considered as a continuous or discrete variable. Here we will compromise and introduce two models, one in the module proper and another in a project. Most of the module is devoted to a discussion of a continuous time model first introduced by A. J. Lotka and subsequently refined by many researchers. The Lotka model has served as the taking off point for much of the research in population dynamics. The topic of one of the projects is a discrete time model due to P. H. Leslie. Both models are discussed in N. Keyfitz's book Introduction to the Mathematics of Population [K1]. This book, the article [K2], and especially the book [KF] contain data and applications of conclusions based on the model to problems in economics, education, social services, etc. In addition, the book [KF] contains computer programs for many of the calculations. There are a few references provided at the end of the module. The items [H] and [K2] contain other references to the vast literature on the subject.

2. Formulation of a Model

In this section we will formulate a continuous time model for the dynamics of the female population. The dynamics of the entire population can be determined from a knowledge of the dynamics of the females, assumptions relating the number of male and female births and assumptions regarding male survivorship. We begin with some definitions and notation.

Let $B(t)$ denote the rate of additions to the female population, the birth rate, at time t. Our first goal will be to relate B to three quantities which can be determined from available data. Specifically, we suppose that the age distribution of the initial population, the age specific mortality rate, and the age specific fertility rate are known quantities. Suppose that we identify $t = 0$ as the time at which our analysis of the population begins, and let $u_0(a)$ denote the density of the initial age distribution. That is, we assume that the number of individuals with ages between a_1 and a_2 at $t = 0$ is given by $\int_{a_1}^{a_2} u_0(a)da$. The mortality rate will be defined indirectly by a function $\ell(a)$ which gives the proportion of births which survive to age a. The function ℓ is usually given as a life table, a listing of age a and survivor ratio $\ell(a)$, which is obtained from census data for the population. A sample life table is given in Table 1.

A Typical Life Table

a age	$\ell(a)$ Proportion of births surviving to age a
0	1.0
5	0.76
10	0.72
15	0.69
20	0.65
25	0.61
30	0.56
35	0.52
40	0.48
45	0.44
50	0.40
55	0.36
60	0.32
65	0.27
70	0.20
75	0.12
80	0.07
85	0.03
90	0.00

Table 1

The fertility rate is obtained from a table of the number of female births produced by females of various ages. In particular, suppose that $\beta(a)$ is the rate at which female offspring are produced by 100,000 females of age a. Then we define $b(a)$, the fertility rate, by $b(a) = 10^{-5}\beta(a)$. That is, the fertility rate is the average rate at which females of age a bear female offspring. We assume that $b(a) = 0$ for large age a, say $b(a) = 0$ for $a \geqq A$. In human populations it is customary to assume $b(a) = 0$ for $a \leqq 15$ and $a \geqq 55$. Notice that we are assuming that the fertility and mortality rates, though age dependent, are independent of time t. That is, ℓ and b depend only on a, not on t.

Next we turn to the derivation of an expression for the birth rate $B(t)$ at time t. These births can (potentially) arise from two sources. First, they can be births to females in the initial population, and second they can be births to females born since $t = 0$. In the first case the density of females surviving from the initial population to attain age a at time t, $a \geqq t$, is

$$[\ell(a)/\ell(a-t)]u_0(a-t).$$

Indeed, $\ell(a)/\ell(a-t)$ is the proportion of the initial population which will survive from age $a - t$ to age a, and $u_0(a-t)$ is the density of females in the initial population which, if they survived, could be of age a at time t. Each of these females will bear female offspring at the rate $b(a)$. Therefore

$$b(a)[\ell(a)/\ell(a-t)]u_0(a-t)$$

is the density function for the rate at which female offspring will be produced by the survivors of the initial population who have reached age a at time t. The contribution of individuals at all ages can be obtained by integrating, and consequently the total birthrate at time t due to the initial population is

$$\int_t^\infty b(a)[\ell(a)/\ell(a-t)]u_0(a-t)da.$$

Since $b(a) = 0$ for $a \geqq A$ this integral is equal to

$$\begin{cases} \int_t^A b(a)[\ell(a)/\ell(a-t)]u_0(a-t)da, & \text{if } t < A \\ 0 & \text{if } t \geqq A. \end{cases}$$

The second source of new females in the population is births to those born after $t = 0$. At time t consider the females of age a. These females were born at the rate $B(t-a)$ at time $t-a$ and the proportion $\ell(a)$ of them survived to age a. Therefore, the number of females of age a to $a+da$ at time t is $\ell(a)B(t-a)da$. Since each of them will bear children at the rate $b(a)$, the rate of additions to the population at time t due to new females is

$$\int_0^t b(a)\ell(a)B(t-a)da.$$

Combining these results we obtain an equation for B:

$$B(t) = \int_t^\infty b(a)[\ell(a)/\ell(a-t)]u_0(a-t)da + \int_0^t b(a)\ell(a)B(t-a)da,$$

which we write as

$$B(t) = h(t) + \int_0^t b(a)\ell(a)B(t-a)da. \qquad (1)$$

This integral equation for the birth rate B is called the renewal equation since it describes the way the population reproduces itself.

If we could solve the equation (1) for the function B, then we could predict the size and age structure of the population; in fact, if $f(a,t)$ denotes the age density of females of age a at time t so that the number of females of age a_1 to a_2 at time t is given by $\int_{a_1}^{a_2} f(a,t)da$, then

$$f(a,t) = \begin{cases} \ell(a)B(t-a), & a < t \\ [\ell(a)/\ell(a-t)]u_0(a-t), & t < a. \end{cases}$$

3. A Stable Age Distribution

Unfortunately, the renewal equation (1) is not an easy equation to solve. It is fairly straightforward using the method of successive approximations to show that under quite general conditions on the functions h, b and ℓ in (1) the renewal equation has a unique solution on $(0,\infty)$ (see for example [BC], p. 217). However, we are interested in certain properties of the solution, in particular its asymptotic behaviour, which are more difficult to obtain. Most of the facts of interest to us, which we will use without proof, can be deduced with the use of Laplace transform methods (see [BC], p. 231 ff).

First, it is known that the solution B has the form

$$B(t) = B_\infty e^{pt} + R(t) \tag{2}$$

where B_∞ and p are constants and the remainder R is a complicated function which satisfies

$$\lim_{t\to\infty} e^{-pt}R(t) = 0.$$

Consequently, R is negligible compared to the first term on the right hand side of (2) for large values of t.

It is easy to determine what the value of p must be. From (1) we have

$$e^{-pt}B(t) = e^{-pt}h(t) + \int_0^t b(a)\ell(a)e^{-pa}e^{-p(t-a)}B(t-a)da.$$

Taking the limit $t \to \infty$, we have

$$B_\infty = B_\infty \int_0^\infty b(a)\ell(a)e^{-pa}da + \lim_{t\to\infty} \int_0^A b(a)\ell(a)e^{-pa}e^{-p(t-a)}R(t-a)da$$

$$= B_\infty \int_0^\infty b(a)\ell(a)e^{-pa}da.$$

Therefore, p must satisfy the characteristic equation

$$1 = \int_0^\infty b(a)\ell(a)e^{-pa}da. \tag{3}$$

This equation was first derived by Lotka in 1922 ([Lo] p. 340). It has a unique real solution for p, say p^*. The constant p^* is called the intrinsic rate of natural growth or the biotic potential of the population. This constant is analogous to the growth rate in the ordinary exponential or Malthusian growth equation.

If b and ℓ are piecewise continuous functions, then it can be shown that there are infinitely many complex solutions of (3). If we denote them by $\{\varphi_n\}$, then the following facts hold

1) The φ_n occur in conjugate pairs, i.e. if φ_n satisfies (3) then so does $\overline{\varphi}_n$,
2) $\mathrm{Re}\varphi_n < p^*$, for all n,
3) $\mathrm{Re}\varphi_n \to -\infty$ as $n \to \infty$.

A formula can be derived for the constant B_∞. The details, which are somewhat involved, can be found in [BC], p. 232 ff.

$$B_\infty = \frac{\int_0^A h(t)e^{-p^*t}dt}{\int_0^A a\ell(a)b(a)e^{-p^*a}da} \tag{4}$$

Now we can show that the population approaches a stable age distribution. To this end, consider

$$\lim_{t\to\infty} e^{-p^*t}f(a,t).$$

From the formula for $f(a,t)$, we see that

$$\lim_{t\to\infty} e^{-p^*t}f(a,t) = B_\infty \ell(a)e^{-p^*a}$$

where B_∞ is the constant determined by (4). Therefore, for large times the population is approximately described by

$$e^{p^*t}B_\alpha \ell(a)e^{-p^*a}. \tag{5}$$

If $p^* > 0$ the population increases with passing time, if $p^* < 0$ it declines and approaches zero. The expression (5) is known as the stable age distribution associated with the population. Notice that the time variation of the stable age distribution is determined by the real solution of the characteristic equation.

One of the consequences of the existence of a stable age distribution is that for large times, the proportion of the population in any age group remains constant. For example, the proportion of the population which is in the age bracket (a_1,a_2) is

$$\frac{\text{Number with ages } a_1 < a < a_2}{\text{Total population}} = \frac{\int_{a_1}^{a_2} f(a,t)da}{\int_0^\infty f(a,t)da}$$

which is approximately equal to

$$\frac{\int_{a_1}^{a_2} \ell(a)e^{-p^*a}da}{\int_0^\infty \ell(a)e^{-p^*a}da}$$

for large t.

The fact that the population approaches a stable age distribution is quite important, and many calculations which demographers make are based on the stable age distribution.

Summary

The theoretical results which we have accumulated so far are easily summarized. Lotka's age dependent theory for the evolution of a female population is based on the renewal equation (1). This equation involves two functions which are determinable from data: $h(t)$, the contribution of the initial population to the births at time t, and $m(a) = b(a)\ell(a)$, the net maternity function which measures the viability and fertility of individuals.

If m is a piecewise continuous function which vanishes for large ages (say $m(a) = 0$ for $a \geqq A$), then the population approaches a stable age distribution. In fact, in that case B can be written as

$$B(t) = B_\infty e^{p^*t} + R(t),$$

where p^* is determined from the characteristic equation (3) as the unique real solution of that equation for p. The remainder $R(t)$ is small when compared to e^{p^*t} for large t, and the constant B_∞ is given by the formula (4).

Exercises

1. The characteristic equation also arises in the formal solution of (1) using Laplace transforms. Derive this formal solution and show how the characteristic equation arises.

2. Since p^* is normally quite small, an approximate value for p^* can often be determined by replacing e^{-pa} in (3) by its Taylor expansion about $p = 0$ and then ignoring the terms of order p^3 and higher. Use this method to obtain a formula for an approximate value of p^*.

Project

In this project you are asked to use given fertility rates, life table and initial age distribution to determine the intrinsic growth rate, stable age distribution, and projections of the population distribution at two specified times in the future. We will follow demographic custom and give data over five year intervals. This suggests the approximation of taking the data to be constant over five year intervals. For example, using the data in Table 1, we would define

$$\ell(a) = \begin{cases} .76 & 0 \leqq a < 5 \\ .72 & 5 \leqq a < 10 \\ .69 & 10 \leqq a < 15 \\ .65 & 15 \leqq a < 20. \\ \cdot & \\ \cdot & \\ \cdot & \end{cases}$$

This approximation simplifies the integrals to finite sums.

Suppose that the initial age distribution is given by column 2 of Table II and the fertility rate is given by column 3 of Table II.

a) Determine the intrinsic growth rate. The formula derived in Exercise 2 may be used as a first approximation. (A good problem in numerical analysis is to use Newton's method to solve the characteristic equation for p. Newton's method and other methods are discussed in [IK], Chapter 3).

b) Evaluate the stable age distribution by using equation (4). (First $h(t)$ must be found for each t, and then the integrals in (4) must be evaluated. This is a lengthy but straightforward calculation).

c) Determine $f(a,5)$ and $f(a,10)$.

Age bracket	number of females ($\times 10^3$) in initial population	number of births per female
[0,5)	105	0
[5,10)	100	0
[10,15)	905	.004
[15,20)	805	.298
[20,25)	650	.732
[25,30)	560	.512
[30,35)	565	.297
[35,40)	620	.157
[40,45)	635	.045
[45,50)	580	.002
[50,55)	535	0
[55,60)	465	0
[60,65)	405	0
[65,70)	300	0
[70,75)	210	0
[75,80)	135	0
[80,85)	65	0
[85,90)	15	0

Table II

4. Population Waves

Oscillations are common in natural phenomena. When given an impulse which forces it away from equilibrium, systems as diverse as an elastic spring, a bowl of jello, or water in a pan respond by oscillating for a period of time and then returning to an equilibrium condition. Under certain conditions populations also exhibit oscillations, and these oscillations are the topic of this section. A more detailed discussion, and applications to social and economic questions may be found in [K2].

A. Bernardelli Waves

In some populations, for example insect populations such as 13 and 17-year cicadas, the reproductive period is restricted to a very short time interval. Unfortunately, the analysis outlined above may not be valid in such cases because the maternity function can no longer be thought of as piecewise continuous. This situation will be illustrated by considering an example involving a maternity function of a more general type than that discussed previously.

Suppose that

$$m(a) = \sum_{j=1}^{k} m_j \delta(a-a_j)$$

where the m_j are positive constants and δ denotes the Dirac delta function. (The Dirac delta function is an example of a generalized function. It is interpreted through values of integrals in which it forms a part of the integrand; i.e., if g is a continuous function, then

$$\int_{-\infty}^{\infty} \delta(x-x_0)g(x)dx = g(x_0). \tag{6}$$

It is somewhat involved to describe generalized functions in a rigorous mathematical setting. It is clear that the delta function cannot be described by giving the value $\delta(x)$ for every real x. Indeed, we would have to write $\delta(x) = 0$ for $x \neq 0$ and $\delta(0) = \infty$. Still, the concept is quite simple: the delta function can be thought of as a device which "pops out" the value of an integrand at one point as in the formula (6) and thereby states that the value of the integral depends only on the value of the integrand at one point, $x = x_0$).

This maternity function indicates that individuals reproduce at ages a_j. For simplicity, suppose that $a_j = j\alpha$ where α is the length between breeding times. With this net maternity function the renewal equation becomes

$$B(t) = h(t) + \sum_{j=1}^{k} m_j B(t-j\alpha)H(t-j\alpha)$$

where H is the Heaviside function

$$H(x) = \begin{cases} 0, & x < 0, \\ 1, & x >). \end{cases}$$

Thus, a collection of individuals born at time t will produce m_1 offspring at time $t + \alpha$, m_2 at time $t + 2\alpha$, etc. The characteristic equation corresponding to this choice of net maternity function is

$$1 = \sum_{j=1}^{k} m_j e^{-pj\alpha} \tag{7}$$

Setting $\nu = e^{-p\alpha}$, we obtain

$$1 = \sum_{j=1}^{k} m_j \nu^j. \tag{8}$$

The right side of this equation is simply a polynomial in ν with positive coefficients. Since we are concerned only with $\nu > 0$ (remember $\nu = e^{-p\alpha}$), since the polynomial is zero for $\nu = 0$ and since the first k derivatives of this polynomial are all positive for $\nu = 0$, we conclude that there is a unique solution ν^* of (8) for $\nu > 0$. Therefore

$$p^* = -(1/\alpha)\log \nu^*$$

satisfies (7). However, in contrast to the situation in section 2, the numbers

$$p^* + (2\pi in/\alpha), \qquad n = \pm 1, \pm 2, \ldots,$$

are also solutions of the characteristic equation (7) since $e^{2\pi in} = 1$ for any integer n. Thus, in this case the analogue of the remainder R in (2) is not necessarily negligible in comparison with e^{p^*t}.

Consider, for example, the special case in which $k = 1$. Here, we have a population of individuals who can reproduce only once, when they reach age α. The renewal equation becomes

$$B(t) = \begin{cases} h(t), & 0 \leq t < \alpha, \\ m_1 B(t-\alpha), & \alpha \leq t. \end{cases} \tag{9}$$

This shows that for $0 \leq t < \alpha$

$$B(t + N_\alpha) = m_1^N h(t).$$

If $m_1 = 1$, then each female will exactly reproduce herself when she reaches age α. Thus, one expects the rate of additions to the population to merely repeat h in each time interval of length α. In fact,

$$b(t + N\alpha) = h(t), \qquad 0 \leq t < \alpha,$$

so that the rate of additions is a periodic function of t with period α. If $m_1 < 1 (>1)$, then the rate of additions will oscillate with wave length α, but it will approach zero (∞) as $t \to \infty$.

If the rate of additions is periodic or nearly periodic, then the phenomenon is referred to as a population wave. Note that the wave length (period) of the oscillation in the rate of additions is the same as the age to reproduction. This feature will be observed in a more general setting later.

The example just studied is the analogue for (1) of an observation made by Bernardelli in 1941 (see [Le] p. 200). He used a discrete (matrix) model. In particular, he showed that if the reproduction of the population is focussed in one age bracket, then there is a population wave. Because of this, we refer to these phenomenal oscillations in the rate of additions caused by sharply restricted reproductive periods as Bernardelli Population Waves.

A maternity function like this one is rather extreme. While it may be relevant to some insect populations, it is not typical of most organisms. However, it does provide insight into population waves in more realistic models. We will pursue this point in part B of this section.

Exercises

1. Give a detailed proof that there is a unique solution $\nu^* > 0$ of equation (8).

2. Select several different functions h and graph the resulting function B of equation (9) for the three cases $m_1 > 1$, $m_1 = 1$, and $m_1 < 1$.

Project

Repeat Bernardelli's analysis by formulating a discrete model for the dynamics of the female population. (See [Le] and [Kl], Chapters 2 and 3.) Hint: Let $(n_0, n_1, \ldots, n_N)$ denote the sizes of the populations in the age brackets,

$$[0, a_1), [a_1, a_2), \ldots, [a_N, a_{N+1}),$$

respectively, where the lengths of these intervals are equal to one reproduction interval, hereafter taken to be the unit of time. Let ℓ_i be the proportion of individuals in $[a_i, a_{i+1})$ who survive one unit of time. Finally, let b_i denote the fertility of individuals in $[a_i, a_{i+1})$.

Then, if $(n_0', n_1', \ldots, n_N')$ denotes the population sizes one unit of time later, we have

$$n_0' = b_0\ell_0 n_0 + b_1\ell_1 n_1 + \ldots + b_N\ell_N n_N$$

$$n_1' = \ell_0 n_0,\ n_2' = \ell_1 n_1,\ \ldots,\ n_N' = \ell_{N-1} n_{N-1},$$

and we no longer account for those who were in $[a_N, a_{N+1})$.

These equations can be written more concisely by taking advantage of matrix notation:

$$\begin{bmatrix} n_0' \\ n_1' \\ n_2' \\ \cdot \\ \cdot \\ \cdot \\ n_N' \end{bmatrix} = \begin{bmatrix} b_0\ell_0 & b_1\ell_1 & \cdot & \cdot & b_{N-1}\ell_{N-1} & b_N\ell_N \\ \ell_0 & 0 & \cdot & \cdot & 0 & 0 \\ 0 & \ell_1 & \cdot & \cdot & 0 & 0 \\ & & & & & \\ & & & & & \\ 0 & 0 & & & \ell_{N-1} & 0 \end{bmatrix} \begin{bmatrix} n_0 \\ n_1 \\ n_2 \\ \cdot \\ \cdot \\ \cdot \\ n_N \end{bmatrix}$$

Or, if we set

$$\vec{n} = \begin{bmatrix} n_0 \\ n_1 \\ \cdot \\ \cdot \\ \cdot \\ n_N \end{bmatrix},\ \vec{n}' = \begin{bmatrix} n_0' \\ n_1' \\ \cdot \\ \cdot \\ \cdot \\ n_N' \end{bmatrix} \text{ and } \mathcal{L} = \begin{bmatrix} b_0\ell_0 & \cdots & b_N\ell_N \\ \ell_0 & \cdots & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & \cdots & 0 \end{bmatrix}$$

then we can write the system as $\vec{n}' = \mathcal{L}\vec{n}$. After two units of time we have

$$\vec{n}'' = \mathcal{L}\vec{n}' = \mathcal{L}^2\vec{n}.$$

etc. This discrete version of the Lotka theory is due to P. H. Leslie (see

[Le]) who developed several properties of solutions by analyzing the coefficient matrix of this model.

The case which Bernardelli described is presented here by taking $b_0 = b_1 = \dots = b_{N-1} = 0$ and $b_N > 0$. This requires no substantial analysis, and in fact

$$n_0' = b_N \ell_N n_N, \quad n_0'' = b_N \ell_N n_N' = b_N \ell_N \ell_{N-1} n_{N-1},$$

and so on, until after $N + 1$ units of time.

$$n_0^{(N+1)} = (b_N \ell_N \cdots \ell_0) n_0.$$

This equation is the discrete version of (9). If $b_N \ell_N \cdots \ell_0 = 1$ then each initial cohort exactly reproduces itself, and the rate of additions will be a periodic function.

B. Population Waves Near Bernardelli Waves

Useful insight into population waves can be obtained by considering the Bernardelli waves as a limiting form. Specifically, let us consider a situation in which the net maternity is slightly "smeared out." For example, let

$$m_\Delta(a) = \begin{cases} m/\Delta, & \alpha - \Delta < a < \alpha, \\ 0, & \text{otherwise}, \end{cases}$$

where Δ is a small positive constant. It can be shown that as $\Delta \to 0$ this function behaves increasingly like the generalized function $m\delta(a-\alpha)$. In this case reproduction takes place not only at age α, but in a short time interval preceding age α also. The solution in this case should approach a Bernardelli wave as $\Delta \to 0$.

Before proceeding with the analysis, stop and think about what effect this "smearing out" of m should have on B. Should B still be periodic or nearly periodic? What should be the wavelength?

Proceeding, we find the characteristic equation in this case to be

$$1 = \frac{m}{\Delta} \int_{\alpha-\Delta}^{\alpha} e^{-pa} da = \frac{m}{p\Delta} e^{-p\alpha} [e^{p\Delta} - 1].$$

When $\Delta = 0$ the Bernardelli case results, and we have

$$1 = me^{-p_0\alpha},$$

from which we conclue that

$$p_{o,n} = (1/\alpha)\log m + (2\pi i n/\alpha), \quad n = 0, \pm 1, \pm 2, \dots .$$

We continue in the hope that the solutions of the characteristic equation for $\Delta > 0$ will be near these values in some sense. (This hope is justified by the implicit function theorem.)

Using these values as starting points, let us determine some of the characteristic roots of the full problem $(\Delta > 0)$ by using Taylor's formula: Fix a value for n and look for a solution (characteristic root) of the form

$$p_n = p_{0,n} + \Delta\, p_{1,n} + (\Delta^2/2)p_{2,n} + \dots .$$

Substituting this expression into the characteristic equation, and equating like powers of Δ in the result, we obtain (after a mildly tedious calculation)

$$p_{1,n} = -(p_{0,n}/2\alpha) \quad \text{and} \quad p_{2,n} = c + id$$

where c and d are rather complicated expressions.

If $m = 1$ the calculations are easier. In this case

$$p_{0,n} = 2\pi i n/\alpha, \quad p_{1,n} = -2\pi i n/2\alpha^2, \quad \text{and}$$

$$2p_{2,n} = -(\pi^2 n^2/\alpha^5) + i(\pi n/\alpha^3).$$

Thus

$$p_n = (-\pi^2 n^2/2\alpha^5)\Delta^2 + 2\pi i n[(1/\alpha) - (1/2\alpha^2)\Delta + (1/4\alpha^3)\Delta^2] + E,$$

where the error term E includes all terms in the Taylor series which contain Δ^3 and higher powers of Δ. This expression contains a great deal of information. For the remaining argument we neglect the error term E. In fact, for small values of n the error term is proportional to Δ^3 while for large values of $\pm n$ the real part of p_n is large and negative and consequently plays a dominant role in the solution.

The solution of the renewal equation corresponding to the choice $b(a)\ell(a) = m_\Delta(a)$ has the form

$$B(t) = \sum_{n=-\infty}^{\infty} \beta_n \exp(p_n t), \qquad \bar{p}_n = p_{-n}$$

Using Euler's formula $e^{a+ib} = e^a(\cos b + i \sin b)$, and the fact that B is real (i.e., $\beta_n = \bar{\beta}_{-n}$, the complex conjugate of β_{-n}), we see that B can be written as

$$B(t) = \sum_{n=0}^{\infty} B_n \exp(t \operatorname{Re} p_n)\cos(t \operatorname{Im} p_n),$$

where $B_0 = \beta_0$, $B_n = 2\beta_n$ for $n > 0$ Re p_n denotes the real part of p_n and Im p_n denotes its imaginary part.

This calculation shows that the real part of p_n measures the growth or damping of the contribution of the nth term to $B(t)$ while the imaginary part measures the frequency of oscillations. As noted earlier, Re p_n is large and negative for large values of n, and consequently $\exp(t \operatorname{Re} p_n)$ is much less than one for such values of n and t not near zero.

Therefore, we may write

$$B(t) = B_0\exp(t\mathrm{Re}p_o)\cos(t\mathrm{Im}p_0) + B_1\exp(t\mathrm{Re}p_1)\cos(t\mathrm{Im}p_1) + \dots .$$

Using the value of p_0 and the approximation developed above for p_1 we have

$$\mathrm{Re}p_0 = 0, \quad \mathrm{Im}p_0 = 0, \qquad \text{(remember } m = 1)$$

$$\mathrm{Re}p_1 = -\pi^2\Delta^2/2\alpha^5, \quad \mathrm{Im}p_1 = 2\pi[(1/\alpha) - (1/2\alpha^2)\Delta + (1/4\alpha^3)\Delta^2].$$

Therefore,

$$B(t) = B_0 + B_1\exp[-\pi^2\Delta^2/2\alpha^5]\cos[2\pi t\Big((1/\alpha) - (\Delta/2\alpha^2) + (\Delta^2/4\alpha^3)\Big)] + \dots, \tag{10}$$

which shows clearly that the solution will be very slowly damped (since for Δ near zero $\Delta^2 t$ changes very slowly) and it will have oscillations with frequencies $(1/\alpha) - (\Delta/2\alpha^2) + (\Delta^2/4\alpha^3)$. The coefficient B_1 is given by the formula

$$B_1 = \frac{\Delta\int_0^{\infty} h(t)e^{-p_1 t}dt}{m\int_{\alpha-\Delta}^{\alpha} ae^{-p_1 a}da}$$

As a result, we see that the smearing out of the Bernardelli case leads to a slowly damped rate of additions which exhibits oscillations having wave lengths given approximately by

$$\Big((1/\alpha) - (\Delta/2\alpha^2) + (\Delta^2/4\alpha^3)\Big)^{-1} \sim \alpha + \Delta/2 + \dots . \tag{11}$$

In particular, for times for which $\Delta^2 t$ is much smaller than one, the solution will appear to be periodic, while for large times the rate of additions will approach the constant B_0.

C. Summary and Conclusions

Lotka defined the mean generation time T of a population to be the value such that

$$e^{p^* T} = \int_0^\infty \ell(a) b(a) da$$

The integral on the right hand side of this expression is called the population's net reproduction rate or NRR for short (see [KF]) and it gives the number of female progeny expected to be born to a female just born. In part A of this section we had

$$\text{NRR} = \int_0^\infty m(a) da = \sum_{j=1}^{k} m_j, \quad \text{and} \quad p^* = -(1/\alpha) \log \nu^*$$

where ν^* was defined as the positive root of $\sum_{j=1}^{k} m_j \nu^j = 1$. Thus

$$T = \alpha\left(-\log \sum_{j=1}^{k} m_j\right) / \log \nu^*$$

and we see that T is proportional to α. In fact, when $k = 1$, $T = \alpha$.

In part B of this section we had

$$\text{NRR} = \int_0^\infty m_\Delta(a) da = m \quad \text{and} \quad p^* = (1/\alpha) \log m.$$

Thus

$$T = \alpha.$$

On the other hand, the wavelength was calculated to be α in Part A and $\alpha + (\frac{\Delta}{2}) + \dots$ in Part B. This shows that the length of the population waves is closely related to the mean generation time in the examples that we have considered.

This observation carries over to more general net maternity functions. This is strikingly illustrated in the data presented by Keyfitz and Flieger where the generation time listed for most populations (Table 7, [KF] Chapter 16) is quite near the wave length ($2\pi/y$ in their notation, end of Table 8, [KF], Chapter 16).

This is an entirely reasonable result: A drastic change in the birth rate should be repeated in successive generations, at multiples of one generation time in the future.

Exercises

1. Determine the terms c and d in the representation $p_{2,n} = c + id$.
2. Determine the coefficient of Δ^2 in the expansion (11).
3. Plot the function B given by (10) for $\alpha = 1$, $B_0 = 1$, $B_1 = 1$, $\Delta = .01$.

Project

As we have seen, oscillations in populations take place over time intervals that are proportional to the mean generation time. Therefore, to observe the oscillations in the population projects, several projections (typically 30 years in length for human populations) must be calculated. This is a major undertaking.

In order to illustrate this oscillatory phenomenon in populations while keeping the calculations at a reasonably simple level, let us consider an imaginary insect population. In particular, let us consider the (as yet undiscovered) 5-year cicadas. These live as adults for only four to six weeks, after remaining in a non-reproductive state for 5 years. Therefore, the analysis outlined in part B of section 4 is applicable, and we take $\alpha = 5$, $\Delta = .15$. Let the contribution to births by the initial population be described by

$$h(t) = \begin{cases} F/\Delta, & \text{for } 5 - \Delta \le t \le 5 \\ 0 & \text{otherwise.} \end{cases}$$

Finally, suppose that each initial cohort will exactly reproduce itself 5 years later, i.e., $m = 1$. It is estimated that the species has been in existence in its present form for 1500 generations.

The aim of this project is to describe the dynamics of the species over 1500 generations. This can be accomplished by evaluating the terms in the expression for B:

$$B(t) = B_0 + \sum_{n=1}^{\infty} B_n \exp(-\pi^2 n^2 \Delta^2 t / 2\alpha^5) \cos(2\pi n t(\frac{1}{\alpha} - \Delta\varphi)),$$

where $\varphi = (1/2\alpha^2) - (\Delta/4\alpha^3)$. How many terms of the series should be retained? We could use the first two terms, but since for $n = 2$ and $t = 5 \times 1500$(years), we have the exponent approximately equal to -1.06, it follows that the second term contains a factor e^{-1}. We view this as negligible.

Note: If we were to consider an actual case, we could study the 17-year cicada for which $\alpha = 17$, $\Delta = 1/170$. In this case we would have to retain many terms in the series to achieve comparable accuracy.

Returning to the case of 5-year cicadas, we will analyze the function

$$B(t) = B_0 + B_1 \exp(-\pi^2\Delta^2 t/2\alpha^5)\cos[2\pi t(\frac{1}{\alpha} - \Delta\varphi)].$$

a) Determine $\Delta\varphi$ and the coefficients B_0 and B_1.

b) Study the emergence at times $t = 5N$ for $N = 1, \ldots, 1500$:

$$B(5N) = B_0 + B_1 \exp[-\pi^2 N(1.8 \times 10^{-5})]\cos[2\pi N(1-5\Delta\varphi)].$$

c) Study the damping. Is it significant? What is the percentage of reduction after 1500 generations? How long will it take for the oscillations to die out? (Remember that we have ignored terms of order e^{-1} in the series expansion of B).

d) Study the periodicity of the solution. Is the solution nearly periodic during the first 1500 generations? The answer to this question is no. This model illustrates a "beat" phenomenon.

To see this, observe that

$$\cos[2\pi N(1-5\Delta p)] = \cos 2\pi N \times \cos(10N\pi\Delta\varphi) = \cos(10N\pi\Delta\varphi)$$

and evaluate this function for $N = 1, 2, \ldots,$ and note that it assumes many values between -1 and $+1$.

There are many other questions of interest. Mainly, features not accounted for in the model should be discussed:

What effect does a variation in the environment have on the population?

How would predation on the population during the reproductive period affect the population's dynamics?

Certainly effects such as these are significant for the population. These are discussed in [HK].

REFERENCES

[BC] Bellman, R. and K. L. Cooke. Differential-Difference Equations. New York, Academic Press, 1963.

[H] Hoppensteadt, F. Mathematical Theories of Populations: Demographics, Genetics and Epidemics. CBMS vol. 20. Philadelphia, Society for Industrial and Applied Mathematics, 1975.

[HK] Hoppensteadt, F. and J. B. Keller. "Synchronization of periodical cicada emergences." (In press).

[IK] Isaacson, E. and H. B. Keller. Analysis of Numerical Methods. New York, John Wiley and Sons, Inc., 1966.

[K1] Keyfitz, N. Introduction to the Mathematics of Population. Reading, Massachusetts, Addison-Wesley Publishing Company, 1968.

[K2] Keyfitz, N. "Population Waves" in T. N. E. Greville ed. Population Dynamics. New York, Academic Press, 1972.

[KF] Keyfitz, N. and W. Flieger. Population: Facts and Methods of Demography. San Francisco, W. H. Freeman Company, 1971.

[Le] Leslie, P. H. "On the Use of Matrices in Certain Population Mathematics." Biometrika, 33 (1945), 183-212.

[Lo] Lotka, A. J. "The Stability of the Normal Age Distribution," Proc. Nat. Acad. Sci., 8 (1922), 339-345.

Chapter 8

MACDONALD'S WORK ON HELMINTH INFECTIONS

Donald Ludwig
University of British Columbia
and
Benjamin D. Haytock
Allegheny College

Introduction

Parasitic infections are among the most important world health problems. Malaria leads the list, followed by schistosomiasis and other helminth infections. These diseases are endemic in most of the tropical and subtropical areas of the world, and their effect can be so severe as to make certain areas nearly uninhabitable. Although large scale efforts have been made to control schistosomiasis, the long-term effects of these efforts have been disappointing. The reasons for this lack of success and suggestions for improved control measures are given by George MacDonald [3].*

MacDonald's results are derived with the aid of a mathematical model which describes the complicated interaction between parasitic worms, human hosts, and secondary hosts (snails). The conclusions MacDonald draws from his analysis of the model are given on pp. 501-502 of [3]. Basically he concludes that a control effort will have little effect unless it is extremely intensive. Once control efforts reach a certain level, that is once a "break point" is reached, then the level of infection will decline sharply, and the disease may be eradicated within a few years. However, before the "break point" is reached, only a small effect will be observed.

The sharpness of the decline in the parasite population at the break point is due to a peculiar feature of helminth infections: in order for the parasites to reproduce, male and female worms must be paired within the body of a human host. As the mean number of worms per human declines, there is also a decline in the probability that a given female will be able to find a male with whom to mate. Therefore once low average numbers of worms per human are reached, a sharp decline in the parasite population will be observed.

Although MacDonald argues eloquently on the need for a mathematical model and the relevance of its conclusions, he does not supply all of the equations which he used in deriving his results. Also, he failed to utilize the fact that certain of his results can be given explicitly in terms of Bessel functions. The following discussion is intended as an introduction to the mathematical features of MacDonald's model and its solution. It is provided to make the paper more accessible to those with a mathematical background.

---

* A copy of this article, reproduced with permission, appears at the end of this module.

The Schistosomiasis Model

Schistosomes have a complicated life history. Male and female worms must mate within a human host. Thereafter, the female deposits eggs, some of which leave the host in the feces. Other eggs may be trapped in various organs of the body, and these cause the symptoms of the disease. If an egg comes into contact with fresh water, a larva called a miricidium hatches. In order to continue the cycle, the miricidium must penetrate the body of a snail. Once a snail is so infected, a large number of larvae called circariae are produced by asexual means. Each circaria swims freely. It seeks to penetrate the skin of a human, in order to complete the cycle.

The main quantities of interest are the mean number of worms infecting each human (denoted by m) and the number of infected snails (denoted by I). We first consider the variation of I with time. If we assume that the relative death rate of infected snails is a constant, which we denote by δ, and that the rate at which snails are infected is proportional to the number of healthy snails, then I will satisfy an equation of the form

$$\frac{dI}{dt} = -\delta I + C(m)(S - I). \tag{1}$$

Here S denotes the total number of snails, and $C(m)$ is the infection rate. This infection rate depends upon the mean number of paired worms per human $P(m)$, since the number of eggs produced (and hence the number of miricidia) is proportional to $P(m)$. Therefore, we may write

$$C(m) = P(m)\,B, \tag{2}$$

where B is a factor which is independent of m. For simplicity, it is also assumed that S is constant.

A similar argument leads to an equation for the rate of change of m. We may write

$$\frac{dm}{dt} = -rm + A\frac{I}{S}\,. \tag{3}$$

Here r denotes the death rate for worms inside a human host, and A is a factor which takes into account the number of circariae produced per infected snail.

The basic equations which govern the course of the infection are (1) and (3). In order to simplify their analysis, we observe that m changes very slowly since the life span of a worm inside a human is on the order of three years. Therefore, we first think of $C(m)$ as constant in equation (1). Let I^* denote the point where $\frac{dI}{dt} = 0$; thus

$$I^* = \frac{C}{C + \delta} S. \tag{4}$$

If $I < I^*$, then $\frac{dI}{dt} > 0$, and I will increase. Likewise, I will decrease if $I > I^*$. We conclude that $I \to I^*$ as $t \to \infty$. This fact can also be verified from the exact solution of (1). MacDonald assumed that if m varies slowly as a function of t, I should be close to I^* (this assumption is discussed in the appendix). Therefore, to a good approximation, we may replace (1) by

$$\frac{I}{S} \cong \frac{I^*}{S} = \frac{B\ P(m)}{B\ P(m) + \delta}. \tag{5}$$

This relation is the same as expression (12) of MacDonald's paper (identify δ with his $-\log p$ and $P(m)$ with his $m\alpha$).

After substitution (5) into (3), we obtain

$$\frac{dm}{dt} = -rm + \frac{A\ B\ P(m)}{B\ P(m) + \delta}. \tag{6}$$

The Poisson Probability Distribution

In order to proceed further it is necessary to obtain an expression for $P(m)$; and in order to do this we must know something about the probability distribution for the worms. We shall think of the population as being one large "region" in space and make the following assumptions about the distribution of worms within this region.

A1. The number of worms in a region is independent of the number of worms in any non-overlapping region.

A2. The probability distribution for the number of worms in any region is the same as that for any other region of the same size. The size of a region is an appropriate measure of its length, area or volume depending upon the physical nature of the region. We suppose that size is measured by one real variable.

A3. If the size of a region is small, then the probability of exactly one worm in that region is approximately proportional to the size of the region.

A4. If the size of a region is small, then the probability of more than one worm in that region is negligeable in comparison to the probability of one worm.

Let $p_n(t)$ denote the probability of n worms in a region of size t. Assumption A2 makes this notation meaningful. Assumptions A3 and A4 may now be stated more precisely as follows.*

A3: $p_1(h) = \lambda h + o(h)$, for some $\lambda > 0$,

A4: $\sum_{k=2}^{\infty} p_k(h) = o(h)$.

The function $p_o(t)$ can easily be determined. We proceed in the following way: from the relationships

$$p_o(t + h) = p_o(t)p_o(h),$$

which follows from A1, and

$$p_o(h) = 1 - \lambda h + o(h),$$

which follows from A4 and the definition of a probability distribution, we obtain

$$\frac{p_o(t+h) - p_o(t)}{h} = \frac{-\lambda h p_o(t) + o(h)}{h}$$

and hence, taking the limit as $h \to 0$,

(7) $$p_o'(t) = -\lambda p_o(t).$$

Since a worm occupies a finite amount of physical space the probability of one or more worms being in a region of size zero is zero. Therefore $P_o(o)$ must be one, and it follows that

$$p_o(t) = e^{-\lambda t}.$$

We can determine $p_n(t)$ in much the same way. Using the assumptions as we did above, we obtain

$$p_n(t+h) = p_n(t)p_o(h) + p_{n-1}(t)p_1(h) + p_{n-2}(t)p_2(h) + \dots + p_o(t)p_n(h)$$

$$= p_n(t)p_o(h) + p_{n-1}(t)p_1(h) + o(h)$$

and consequently

---

* $o(h)$ denotes a term $t(h)$ satisfying $\lim_{h \to 0} \frac{t(h)}{h} = 0$.

$$(8)\quad p_n'(t) = \lim_{h\to 0} \frac{p_n(t+h)-p_n(t)}{h} = \lim_{h\to 0} \frac{-\lambda h p_n(t) + \lambda h p_{n-1}(t) + o(h)}{h}$$

$$= -\lambda p_n(t) + \lambda p_{n-1}(t),\quad n = 1, 2, \ldots .$$

The differential equations (8) can be solved successively (along with the conditions $p_n(0) = 0$ for $n \geqq 1$) to get

$$(9)\qquad P_n(t) = e^{-\lambda t}\, \frac{(\lambda t)^n}{n!},\quad n = 1, 2, 3, \ldots .$$

Exercise 1. Justify each of the steps in the derivation of equations (7) and (8) and verify that formula (9) actually provides the solution.

Exercise 2. Show that $\sum_{n=0}^{\infty} p_n(t) = 1$. Why is this to be expected?

If we choose our unit of size in such a way that the average human being has size 1 and let p_n denote the probability of a given person carrying n worms, it follows that $p_n = p_n(1) = \frac{e^{-\lambda}\lambda^n}{n!}$. This probability distribution is known as the Poisson distribution.

Exercise 3. Show that if $p_n = \frac{e^{-\lambda}\lambda^n}{n!}$, then the expected number of worms per person is λ.

Calculation of P(m)

If m is the expected number of parasites harbored by a given human, we shall assume that the expected numbers of male and female parasites are each $\frac{m}{2}$. Moreover, we shall assume that the numbers of male and female parasites are independent random variables each with a Poisson distribution. Therefore

$$(10)\qquad \text{Prob}\,[p \text{ males}] = e^{-\frac{m}{2}}\, \frac{1}{p!}\, \left(\frac{m}{2}\right)^p,$$

$$(11)\qquad \text{Prob}\,[q \text{ females}] = e^{-\frac{m}{2}}\, \frac{1}{q!}\, \left(\frac{m}{2}\right)^q,$$

$$(12)\qquad \text{Prob}\,[p \text{ males and } q \text{ females}] = e^{-m}\, \frac{1}{p!q!}\, \left(\frac{m}{2}\right)^{p+q}.$$

For simplicity, we shall assume that pairs are formed whenever possible, i.e., if there are p males and q females present, the number of pairs is given by $\min(p,q)$. Therefore, the number of paired parasites is given by $2\min(p,q)$. Hence, from (12) we conclude that

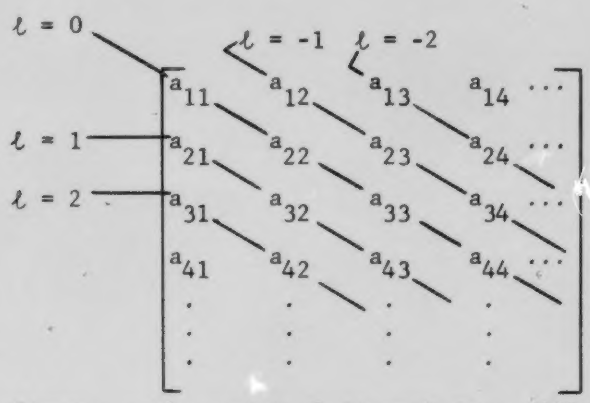
$$P(m) = \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} 2\min(p,q)\, e^{-m} \frac{1}{p!q!} \left(\frac{m}{2}\right)^{p+q}. \tag{13}$$

From formula (13), it is already apparent that

$$P(m) \sim \frac{m^2}{2} \quad \text{for small values of } m, \tag{13'}$$

since the only term which is not of higher degree corresponds to $p = 1$, $q = 1$. For large values of m, one could approximate the Poisson distribution by a normal distribution, and show that $P(m)$ is approximately equal to m. Instead of elaborating upon these remarks, we shall evaluate the sum (13) in terms of known functions. The following calculation is based upon Leyton [2] and Nåsell and Hirsch [4].

The expression (13) will be evaluated by summing the doubly infinite matrix along diagonals $p = q + \ell$, with ℓ fixed. If the matrix is denoted (a_{pq}), then the diagonals are shown in the matrix below



As a preparation, we first sum the original distribution (12) along diagonals. If $\ell > 0$, we define

$$I_{\ell}(m) = \sum_{q=0}^{\infty} \frac{1}{q!(q+\ell)!} \left(\frac{m}{2}\right)^{q + q + \ell}. \tag{14}$$

It turns out that $I_{\ell}(m)$ is a modified Bessel function (see Abromowitz and Stegun [1], page 375). This is a considerable advantage, since these functions have been studied and tabulated. In view of the symmetry of the distribution, we set $I_{-\ell} = I_{\ell}$. By summing over all of the diagonals, we obtain the identity (see [1], page 376, equation 9.6.37)

$$(15) \qquad e^{-m} \sum_{\ell=-\infty}^{\infty} I_{\ell}(m) = 1.$$

This expresses the fact (which can be verified by summing (12) in the usual way) that the probabilities of all of the possibilities add up to 1.

Let A_{ℓ} denote a diagonal sum corresponding to (13); if $\ell \geqq 0$

$$(16) \qquad A_{\ell} = \sum_{q=0}^{\infty} \frac{q}{q!(q+\ell)!} \left(\frac{m}{2}\right)^{2q+\ell}$$

We set $A_{-\ell} = A_{\ell}$, in view of the symmetry in (13). It follows from (13) and (16) that

$$(17) \qquad P(m) = 2e^{-m} \sum_{\ell=-\infty}^{\infty} A_{\ell}.$$

On the other hand a comparison of (14) and (16) shows that if $\ell \geqq 0$, then

$$(18) \qquad A_{\ell} = \frac{m}{2} I_{\ell+1}(m).$$

Therefore (17) can be rewritten as

$$(19) \qquad P(m) = m\, e^{-m}\left(\sum_{\ell=1}^{\infty} I_{\ell} + \sum_{\ell=2}^{\infty} I_{\ell}\right).$$

The identity (15) then implies that

$$(20) \qquad P(m) = m[1 - e^{m}(I_o(m) + I_1(m))].$$

The factor within the brackets gives the proportion of paired parasites. MacDonald denotes this quantity by α and has tabulated it in the second column of Table II, page 492. The asymptotic expansion of 9.7.1 in [1], page 377, shows that if m is large, then

$$(21) \qquad \frac{P(m)}{m} \sim 1 - \left[\frac{2}{\pi m}\right]^{\frac{1}{2}}.$$

This approximation yields $\frac{P(m)}{m} \cong .748$ for $m = 10$. It is even more accurate for larger values of m. For small values of m the table shows that indeed $\frac{P(m)}{m}$ is close to $\frac{m}{2}$.

Exercise 4. Verify the computations in equations (14) through (20).

Analysis of Equation (6)

We begin by looking for equilibrium solutions; i.e., those for which $\frac{dm}{dt} \equiv 0$. Using the notation $\alpha(m)$ for $\frac{P(m)}{m}$ (MacDonald's α) and setting $\frac{dm}{dt} = 0$ in equation (6) we obtain

$$-rm + \frac{AB\ m\ \alpha(m)}{Bm\ \alpha(m) + \delta} = 0. \tag{22}$$

Clearly, $m \equiv 0$ is one solution. To find the others it is helpful to rewrite equation (22) in the form

$$\alpha(m) = \frac{r\delta}{AB - mBr}. \tag{23}$$

The graph for $y = \alpha(m)$ can be constructed from the graphs of $e^{-m}I_o(m)$ and $e^{-m}I_1(m)$ found in [1], page 375. It follows from (13') that the slope of this graph at the origin is $\frac{1}{2}$, and it follows from (21) that the graph is asymptotic to the line $y = 1$. The graph of $y = \frac{r\delta}{AB - mBr}$ is a branch of a hyperbola, and the solutions of (23) correspond to the points of intersection shown in Figure 1.

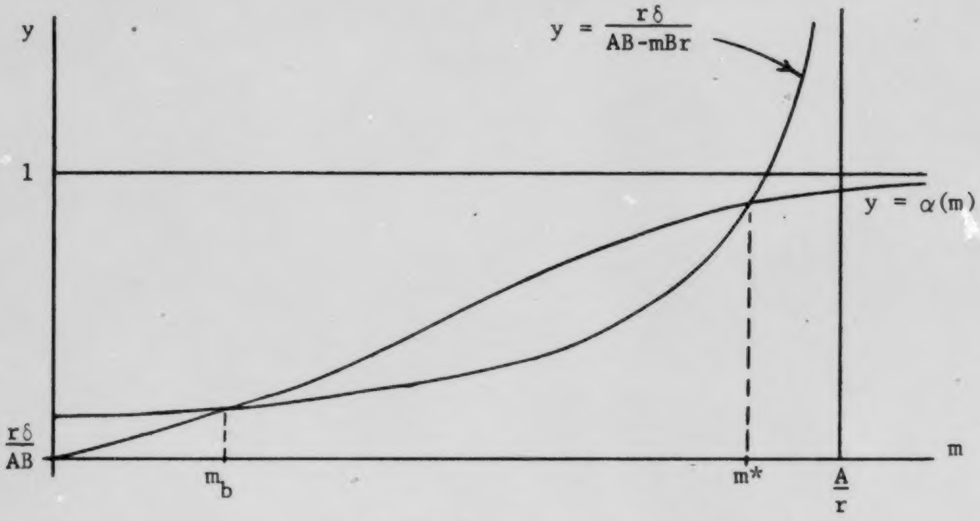


Figure 1

Notice that for the particular choices of the parameters A, B, r, δ corresponding to the graphs shown in Figure 1 there are two positive roots of equation (22), denoted by m^* and m_b. Thus, there are three equilibrium solutions of (6),

$$m \equiv 0$$

$$m \equiv m_b$$

$$m \equiv m^*.$$

Rewriting (6) in the form

$$\frac{dm}{dt} = \left(\frac{m}{mB\alpha+\delta}\right)\left(AB-mBr\right)\left(\alpha - \frac{r\delta}{AB-mBr}\right)$$

it is clear that if $m > m^*$, then $\frac{dm}{dt} < 0$; and if $m_b < m < m^*$, then $\frac{dm}{dt} > 0$. Thus m^* is stable; and if m is any solution of (6) with $m > m_b$, then $m(t) \to m^*$ as $t \to \infty$. If $0 < m < m_b$, then $\frac{dm}{dt} < 0$ and $m \to 0$. The number m_b is the number called the breakpoint. If a treatment program succeeds in bringing m below m_b, then the infection will die out. If not, the level of infection will again rise to m^*.

A comprehensive program aimed at controlling or eradicating the infection will affect all of the parameters A, B, r and δ as well as m. The relationship of the points m_b and m^* to A, B, r and δ lies at the heart of MacDonald's paper. A qualitative feeling for these relationships can be obtained from Figure 2.

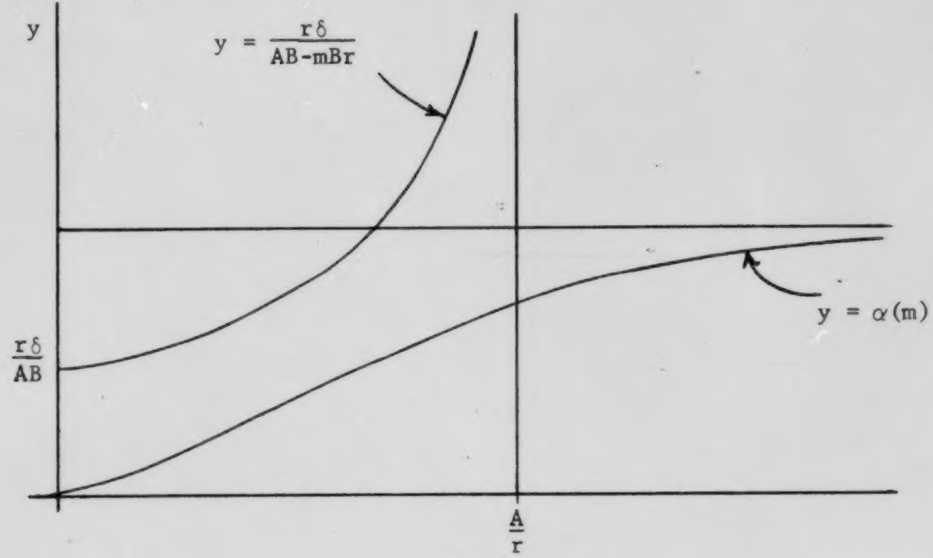


Figure 2

Sanitation improvements have the effect of decreasing B. This in turn has the effect of shifting the hyperbola upward which leads to a shift of m_b to the right. The vertical asymptote at $\frac{A}{r}$ is unaffected by B and consequently large changes in B lead to relatively small changes in m^*. A concentrated program of medical treatment has several effects, one of which is to reduce m and the second of which is an increase in r, the death rate of worms. The vertical asymptote is shifted to the left, the intercept $\frac{r\delta}{AB}$ is raised and the effect is to decrease m^* and increase m_b. A campaign to kill snails has a similar effect since A is decreased and δ increased.

The ideal situation is to modify the parameters in such a way that the hyperbola does not intersect $\alpha(m)$ at all (Figure 2). In this case, the only stable solution is $m \equiv 0$ and every solution will die out.

With this introduction, the reader is urged to turn to MacDonald's original work. A more elaborate and complete mathematical treatment of this problem is contained in Nåsell and Hirsch [4].

Appendix

An Alternate Model

In deriving equation (6) which is the basis of his model MacDonald made the assumption that since m varies slowly with respect to time, I will remain near an equilibrium value I^* (see pages 493-494 of MacDonald's paper or page of this introduction). The assumption seems somewhat dubious however when one considers that killing snails is one of the major methods of control. One might therefore expect equations (1) and (3) to provide a more realistic model. In what follows, we will sketch a phase portrait for the system of equations (1), (3). We will see that from a qualitative point of view we can draw conclusions similar to MacDonald's but that the interdependence of m and I will become more apparent.

The equations under consideration are

$$\frac{dI}{dt} = -\delta I + B\,P(m)(S - I) \tag{1}$$

$$\frac{dm}{dt} = -rm + \frac{A}{S} I . \tag{3}$$

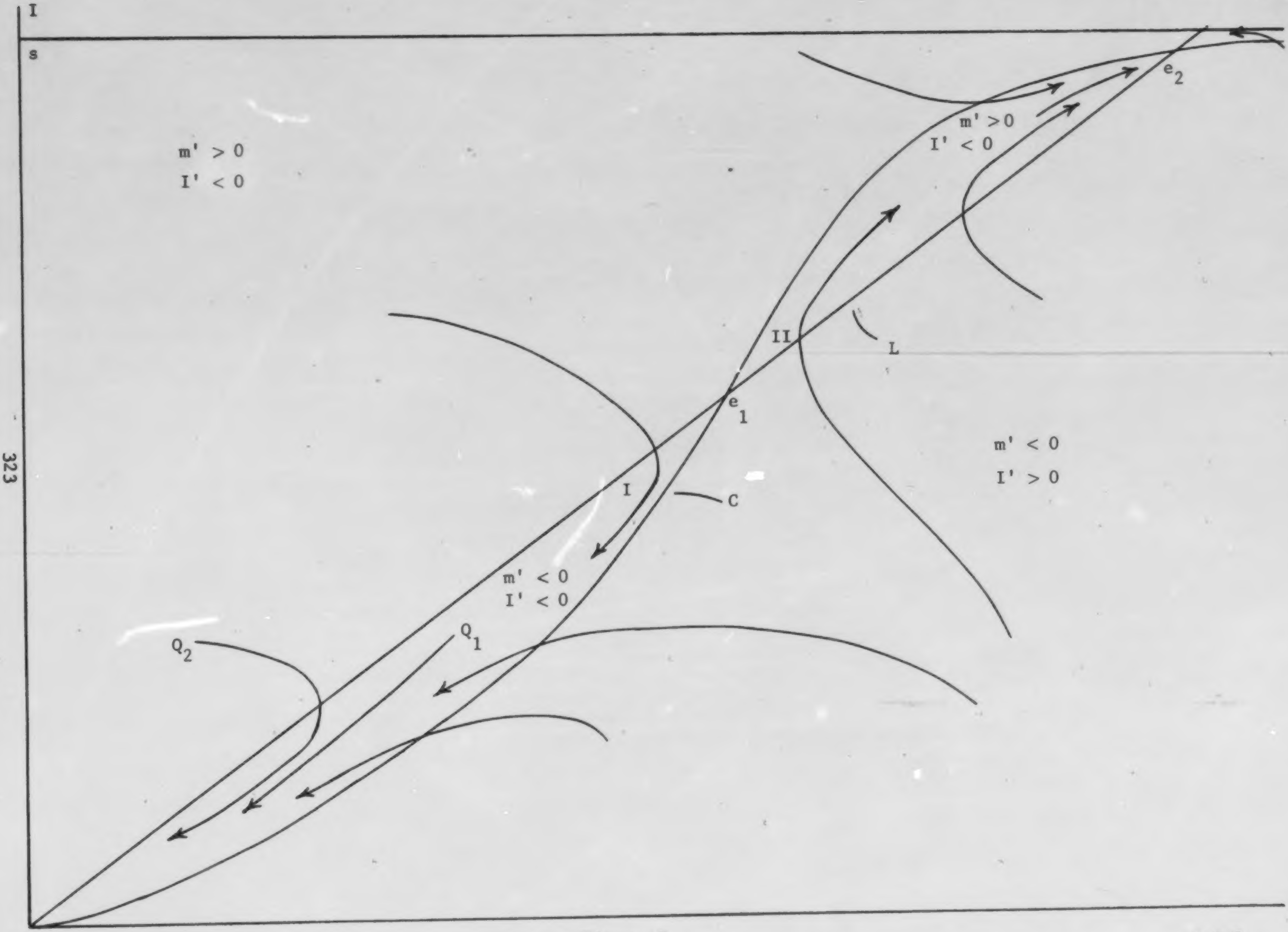

I
s
m' > 0
I' < 0
m' >0
I' < 0
e2
II
L
e1
m' < 0
I' > 0
I
C
m' < 0
I' < 0
Q1
Q2

Figure 3

We begin by observing that $\frac{dI}{dt} = 0$ whenever $\delta I = B\ P(m)(S - 1)$ or when $I = \frac{B\ P(m)S}{B\ P(m) + \delta}$. This curve, labeled c, is sketched in Figure 3. Since $P(m) \sim m$ for large values of m, c will be asymptotic to the line $I = S$. For small values of m, c behaves like the parabola $I = \frac{m^2 S}{2\delta}$. Above c, $\frac{dI}{dt}$ is negative and below c, $\frac{dI}{dt}$ is positive.

Similarly, $\frac{dm}{dt} = 0$ when $I = \frac{S}{A} rm$. This line is labeled L. Above L $\frac{dm}{dt}$ is positive and below L, $\frac{dm}{dt}$ is negative. The three points of intersection 0, e_1 and e_2 are the equilibrium solutions.

Now consider a trajectory starting at a point Q_1 inside region I. Such a curve must move toward the origin since both $\frac{dI}{dt}$ and $\frac{dm}{dt}$ are negative. Furthermore, such a trajectory cannot cross either line L or the curve c (Why?) and so must remain in region I and continue in to 0. Similarly, any trajectory starting in region II must go into the point e_2.

A trajectory starting at a point such as Q_2 must move down and to the right until it crosses L (vertically). Once it enters region I it must behave as in the case above and go into 0.

Exercise 5. Complete the analysis and show that the phase portrait must look like that sketched in Figure 3. Show that 0 and e_2 are both asymptotically stable while e_1 is unstable.

Exercise 6. Discuss how the phase portrait changes as the parameters A, B, r, δ and S are varied. Recall that a change in S also affects A.

Exercise 7. Discuss the similarities and differences between these results and MacDonald's.

References

[1] M. Abromowitz and I. A. Stegun. Handbook of Mathematical Functions. Dover Publications, New York, 1958.

[2] M. K. Leyton. Stochastic Models in Populations of Helminthic Parasites in the Definitive Host. II. Sexual Mating Functions. Math. Biosci. 3 (1968), 413-419.

[3] G. MacDonald. The Dynamic of Helminth Infections, with Special Reference to Schistosomes. Trans. Roy. Soc. Trop. Med. and Hyg. 59 (1965), no. 5.

[4] I. Nåsell and W. M. Hirsch. A Mathematical Model of Some Helminth Infections. Comm. Pure Appl. Math. 25 (1972), 459-478.

PAGES 326-343 REMOVED PRIOR TO BEING SHIPPED TO
EDRS FOR FILMING DUE TO COPYRIGHT RESTRICTIONS.

Chapter 9

MODELING LINEAR SYSTEMS BY FREQUENCY RESPONSE METHODS

William F. Powers
Department of Electrical Engineering
The University of Michigan

PREREQUISITES: Calculus, Differential Equations

NUMBER OF STUDENTS ON PROJECT: Although the project could be done alone, it is recommended that either two or three students work together.

SUBROUTINES (Optional): IBM Scientific Subroutines for Function Minimization (e.g., FMCG, FMFP, NEWT).

POTENTIAL LOCAL EXPERTS: Faculty in engineering.

PREFACE

From a scientific point of view this module is concerned with illustrating some of the problems that arise in the development of models of the performance of human subjects on certain tasks, such as steering. In particular, it demonstrates the use of frequency response techniques in mathematical modeling. Although the module is intended as a case study in modeling, Section 3 is basically self contained and may be appropriate for use in a course on ordinary differential equations.

A brief outline of the various sections is as follows:

Section 1: This section describes the basic type of human behavior which is to be modeled, and previous applications of these models are noted. This section is mainly descriptive.

Section 2: This secion discusses some of the problems involved in obtaining data from experiments involving human subjects and the basic experimental setup. The motivation for employing frequency response mathematical methods is also presented. Again, this section is mainly descriptive.

Section 3: In this section a specific mathematical model is formulated and analyzed. Important results are contained in Theorem 1 (especially part c), Theorem 2, and the corollaries. Both the mathematical and the intuitive aspects of these results are helpful in understanding the model. It may be worthwhile for the instructor to familiarize the students with the basic aspects of frequency response plots beforehand. The manner in which §3.6 is handled will depend upon the time available; the goal is to demonstrate the roles of least squares and function minimization techniques in parameter estimation. In §3.7 the very difficult problem of model verification is considered. In addition to the specific technical questions which arise in verifying the model discussed in this module, one can raise general philosophical questions about the contribution of mathematical modeling at this point.

Section 4: This section contains data from actual human operator experiments. The instructor may use the data in a variety of ways, e.g., (1) give it to students with only an explanation of the experimental procedure and have them hypothesize the model, estimate the parameters of the model, and present arguments for and against the model with respect to intuition; (2) give the students the two models (Eqs. 16, 17) along with the data and have them discuss the trends of the parameters with increasing bandwidth, and determine the physical basis for each component in the models (this approach would not require any numerical procedures).

Section 5: This section briefly discusses current research areas and provides examples of references of recent developments. (Such information would be of primary interest to students who wish to pursue the subject further.)

1. INTRODUCTION

In his classic book Cybernetics [1] Norbert Wiener viewed the human being from an information processing and control system point of view. Such a viewpoint has led to the development of communications and control system oriented mathematical models for human reactions and for physiological components of the human body. Much of this work has involved frequency response techniques, and this module will be concerned with the use of such techniques in mathematical modeling.

Briefly, frequency response techniques are based on the Fourier representations of functions. These representations describe a function in terms of the "amounts" of components with various frequencies that combine to produce the function. For example, if $f(t)$ has a Fourier series expansion

$$f(t) = \sum a_n \sin n t,$$

the different sine functions represent different frequencies and the coefficients are the "weights" with which they are combined. (Non periodic functions f can be represented by Fourier integrals like

$$f(t) = \int a(x) \sin x t \, dx$$

employing a continuum of frequencies, weighted by $a(x)$). Thus the dynamics of certain systems can be recovered from knowledge of their responses to sinusoidal inputs.

The major emphasis of the module is on modeling the human as an operator in a "compensatory task." More specifically, a mathematical model for the human operator is developed from data obtained from experiments in which the subject is supposed to null (i.e. to eliminate) a visual error signal by manipulating a control system with known control dynamics. The visual error signal is random or "random appearing." An example of such a situation is the steering of an automobile simulator. That is, suppose the difference between the center-line of the car and a random (or random appearing) road-marker-line (the input) is displayed to the driver. This will cause the driver to turn the steering wheel, which in turn affects the path of the automobile and in turn the error signal. The dynamic characterists of the system that the subject controls are assumed known; the automobile's steering system in our example. A "feedback" diagram of this situation is shown in Figure 1.

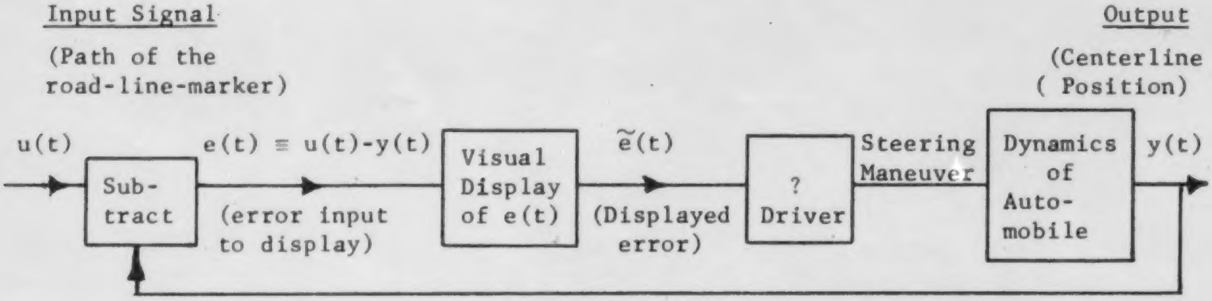

Input Signal
(Path of the road-line-marker)
Output
(Centerline (Position)
u(t)
Sub-tract
e(t) ≡ u(t)-y(t)
(error input to display)
Visual Display of e(t)
ẽ(t)
(Displayed error)
? Driver
Steering Maneuver
Dynamics of Auto-mobile
y(t)

Figure 1. Experiment for Mathematical Modeling of an Automobile Driver.

Approximate mathematical models can then be developed for the driver by frequency response analysis of error signal (programmed path for the road-marker-line minus the actual path) and the output (the resultant center-line position of the automobile). That is, given measurements of the error viewed by the driver, $\tilde{e}(t)$, and the output, $y(t)$, one then applies frequency response methods to those signals to determine a model of the combined driver-vehicle system (see Figure 2). Since the mathematical model of the vehicle is known,

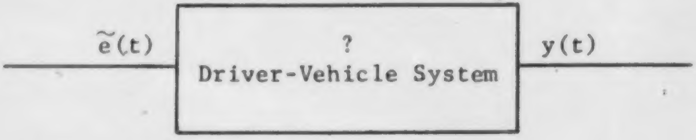

ẽ(t)
?
Driver-Vehicle System
y(t)

Figure 2. Input-Output System for Analysis.

the mathematical model of the driver may then be extracted from the combined model.

The type of modeling discussed above is an active research area, and a number of models developed by this approach have been used in applications. For example, such analyses have been used to show that: the Saturn-Apollo space booster can be flown manually by the astronauts in a contingency situation [2], larger than currently permitted trucks and buses on highways may be uncontrollable in certain situations because of human operator limitations [3], and inclusion of pilot models in aircraft design improves the resultant aircraft performance [4]. In addition, this modeling technique has been employed in the generation of numerous physiological models, some of which are surveyed in references [5], [6], and [7].

The remainder of the module is as follows: in Section 2 the experimental aspects of the human operator modeling problem are discussed along with the motivation for the utilization of frequency response techniques; in Section 3 a tutorial modeling problem which demonstrates the basic mathematics is presented (if desired, the student may perform his own experiment and modeling problem); in Section 4 data from actual human operator experiments are presented for use in modeling and interpretation by the student; finally, in Section 5 extensions which involve more sophisticated mathematics are discussed briefly along with appropriate references to the literature.

2. MODELING THE HUMAN OPERATOR

Because of the large degree of variability between humans, one would not, at first glance, expect to be able to form a mathematical model for the behavior of the human. However, for certain well-defined tasks which involve a human operator, good approximate mathematical models for the nominal (normal) operator of the task may be developed. Such models are then useful in the design of the equipment involved in the particular task. In this section we shall discuss some of the basic considerations involved in the development of human operator models for compensatory tasks (i.e., error nullification).

Assuming a well-defined task for the operator, one must devise an experiment to generate data useful for the development of the model. In Figure 3 a typical laboratory setup of an experiment involving the modeling of a compensatory task is shown. The human operator will move the stick in an attempt to null out the time-varying error signal displayed on the screen. A computer generates the error signal and also contains the dynamics of the physical system to be controlled (e.g., automobile, aircraft).

Since such an experiment involves human subjects a number of considerations must be taken into account. Some of the major considerations are:

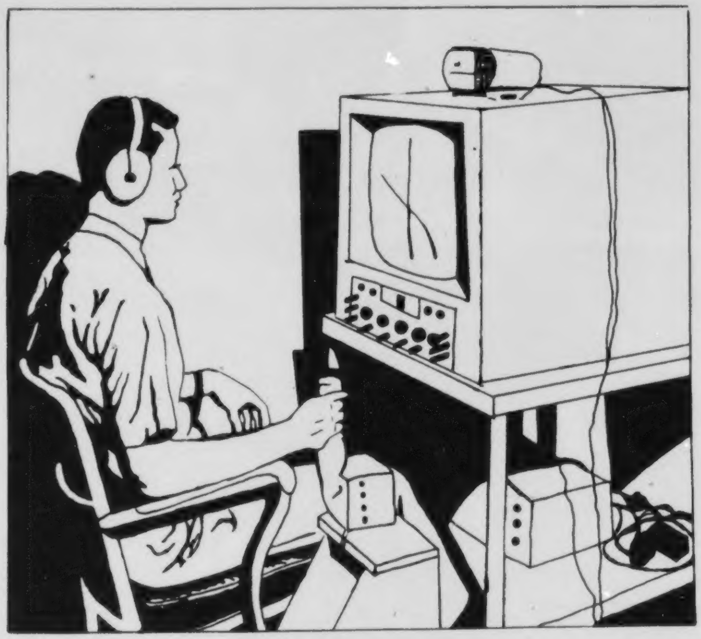


Figure 3. Typical Experimental Setup

(1) Motivation: the human subjects involved in the experiment should be motivated to approximately the same degree to perform to the best of their ability during the test.

(2) Learning: typically the longer one does the experiment, the better the performance (up to a certain point). Such a period is a learning period which should not be part of the data track. Thus, the subject usually does a number of trial experiments before data are collected.

(3) Training: some subjects may be better at performing a compensatory task than other subjects, e.g., a pilot is usually more skilled than a subject who is not a pilot. Such information should be noted since the parameters of the model will be affected.

(4) Attention Span: the experiment should be long enough to get a good data track, but not so long that the subject loses interest or full attention. Typical experiments (such as Figure 3) are on the order of two to five minutes, of which the initial and terminal portions are usually deleted because of learning and decreasing attention, respectively.

(5) Physical and Psychological Conditions: the general physical and psychological condition of the subject should be noted. (In this respect such experiments can be used to determine the effects of, for example, drugs on the ability to do a compensatory task.)

(6) Input Signals: the displayed error signal should not be so simple that the subject can predict it from instant-to-instant, and yet not so complicated that mathematical analysis is impossible.

The last consideration noted above, i.e., the mathematical character of the input signals, is critical since it will strongly influence the type of mathematics to be employed in the data analysis and model building. Let us consider some typical inputs and hypothesize the ability of the subject to follow them. If a constant error signal is displayed, the normal subject should have little trouble following it. Note that a constant signal is a sinusoidal signal with frequency equal zero (i.e., $c = c \cos \omega t$ with $\omega = 0$). Next, suppose a sinusoidal signal of very low frequency is displayed. Again, the subject should be able to track the error signal. However, as the frequency of the error signal increases, the tracking error should increase and finally there will come a point when the subject can no longer track the signal even knowing what the signal is (i.e., a high-frequency sine wave). This frequency is a rough indicator of the "bandwidth" capability of the human operator (typically 5-10 radians/second). Since the human cannot physically track above that bandwidth, input signals containing frequencies higher than the bandwidth are of no use in the development of a mathematical model for the human's behavioral characteristics in tracking an error signal. This argument suggests that one possible classification of input signals is according to frequency. If this classification is adopted, then the admissible frequency range is $[0, BW]$, where $BW \equiv$ bandwidth.

Since frequency is a very important variable in many electrical and mechanical systems, there exist numerous mathematical techniques and theories associated with the frequency characteristics of systems. Such approaches are usually referred to as "frequency domain" approaches.

If the system to be modeled is approximately linear, then there exist a number of frequency domain results which can be applied directly in its identification and modeling. However, one would expect that a human operator does not behave in a linear manner. In fact, one of the pioneers of human operator modeling, A. Tustin, used a simple experiment to show that the human operator is nonlinear. As will be shown in the next section, if a system is described by a linear differential equation with a sinusoidal forcing function of frequency ω, then certain interesting solutions (steady state solutions) can contain sinusoidal components only of frequency ω. Tustin [8] employed an input of three sine waves of different frequencies in a human operator experiment and obtained an output containing frequencies other than the three input frequencies. This result implies that the human operator is nonlinear. However, Tustin also argued that for many tasks the human operator is near-linear and that there exists an approximate linear model which accounts for the dominant features of the human operator's output. Because of this, frequency domain techniques of linear analysis have often been employed in the modeling of human operators.

In the next section basic deterministic techniques associated with frequency domain modeling will be discussed. For a complete discussion of techniques associated with human operator modeling, the theory of stochastic processes is necessary. Indeed, the class of input functions is frequently random, usually a sum of sine functions with random amplitudes and phases; typically ten or more frequencies are employed on the interval $[0,BW]$ so that the frequency is "random appearing." However, even in such cases the final model is often a deterministic one and knowledge of the mathematics in the next section allows a good understanding of the interplay between mathematics and the modeling. Stochastic process techniques are mainly used in the generation of data plots from the experiment and in the interpretation of the nearness-to-linearity of the model.

3. PROPERTIES AND TECHNIQUES OF FREQUENCY RESPONSE MODELING

In this section a simple experiment will be described and used to illustrate mathematical models of the frequency-response type. For students with some background in engineering or physics the governing equations for the motion should be obvious by Newton's laws. However, even for students without such background the experiment still demonstrates a way of "discovering" the laws (or model), and the mathematical ideas behind frequency domain techniques.*

3.1 Description of the experiment.

The following simple experiment will introduce the ideas and mathematical techniques associated with the application of frequency response plots to modeling linear or near-linear dynamic systems.

EXPERIMENT: String rubber bands together to a length of approximately two feet, and tie a weight, e.g., a heavy bolt, to the end of the string of rubber bands. Our purpose is to model this system so that we will be able to predict the motion of the weight, i.e., the output, due to a time-varying hand-motion at point a, i.e., the input.

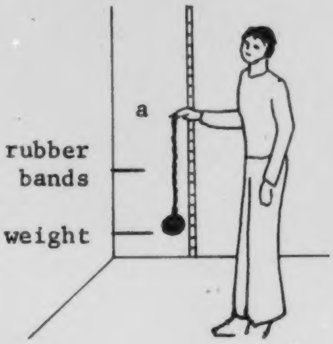


Figure 4. Experimental Setup

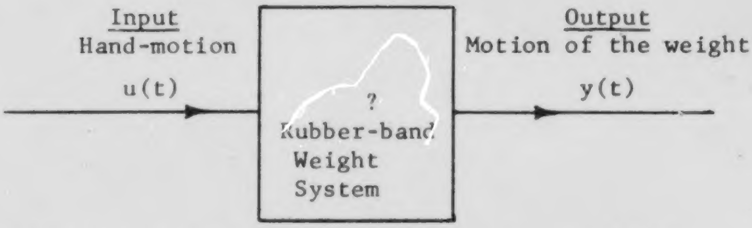


Figure 5. Input-Output Characterization of the Modeling Problem

---

* An alternative to this experiment was devised during the trial teaching of the module. For a description see page 37.

3.2 Models based on linear differential equations.

A mathematical model which provides a useful approximate description of the system discussed above is one involving a linear differential equation with constant coefficients. This choice of a model will be supported below.

The differential equation to be used has the form

$$(1) \qquad \frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \dots + a_1\frac{dy}{dt} + a_0 y = u(t).$$

Such an equation is said to be time invariant since the coefficients on the left are independent of t. The function u is called the forcing function or the input function. A standard problem is to investigate the behavior of the system for various input functions. We shall be concerned in particular with input functions of the form $u(t) = \sin \omega t$ (or with linear combinations of such functions). The number ω is called the frequency of the input; if ω is small, the input is said to be a low frequency input (e.g., 1 cycle/ sec = 2π radians/sec = 1 Hertz); if ω is large the input is said to be a high frequency input (e.g., 10 cycles/sec).

It will be assumed that the reader has a basic working knowledge of the theory of linear differential equations with constant coefficients and with Laplace Transform methods for solving such equations. This material is contained in most textbooks on elementary differential equations, e.g., [20], [21].

Let $y(t)$ be the solution of (1) with $y(0) = y'(0) = \dots = y^{(n-1)}(0) = 0$ and take the Laplace transfor of (1) to obtain the transformation equation

$$(s^n + a_{n-1}s^{n-1} + \dots + a_1 s + a_0)V(s) = W(s)$$

where V and W are the Laplace transforms of y and u respectively. If s is not a root of $s^n + a_{n-1}s^{n-1} + \dots + a_1 s + a_0 = 0$, then

$$(2) \qquad V(s) = \frac{W(s)}{s^n + \dots + a_1 s + a_0}, \quad \text{or}$$

$$(2') \qquad V(s) = W(s)Y(s),$$

with

$$Y(s) = (s^n + \dots + a_1 s + a_0)^{-1}.$$

The function Y is called the transfer function for the equation (or the physical system modeled by the equation). Equation (2') says that the Laplace transform of y may be found by multiplying the Laplace transform of u by the transfer function. The reader should note that the transfer function may be determined from (1) by inspection and that, conversely, the coefficients in (1) can be determined from the transfer function.

We now consider some very important properties of linear, time invariant differential equations.

Definition: The differential equation

$$\frac{d^n y}{dt^n} + a_{n-1}\frac{d^{n-1}y}{dt^{n-1}} + \ldots + a_1\frac{dy}{dt} + a_0 y = 0$$

is said to be stable if all solutions y have the property that $\lim_{t\to\infty} y(t) = 0$.*

Definition: Equation (1) is said to be stable if the corresponding homogeneous equation is stable.

In what follows j is used to denote $\sqrt{-1}$. This notation is common in engineering literature.

Theorem 1. Consider an equation of the form (1) with real coefficients $a_0, \ldots, a_{n-1}$ and with $u(t) = \exp(j\omega t)$. Then

a) (1) is stable if and only if all of the roots $\lambda_1, \ldots, \lambda_n$ of the characteristic equation $s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0 = 0$ have negative real parts.

b) if Y (the transfer function of (1)) is defined at $j\omega$, then the general solution of (1) can be written as $y_g = y_T + y_{ss}$ where y_T is the general solution of the homogeneous equation and y_{ss} is the steady state solution of (1) defined by

$$y_{ss}(t) = Y(j\omega)\exp(j\omega t). \tag{3}$$

If all the λ_k are distinct, then $y_T(t) = \sum_{k=1}^{n} c_k \exp(\lambda_k t)$.

c) if (1) is stable, then $\lim_{t\to\infty} y_T(t) = 0$ and y_g is asymptotic to y_{ss}

The function y_T is called the transient part of y_g.

---

* Some authors say the equation is asymptotically stable if for every solution $\lim_{t\to\infty} y(t) = 0$.

The proof of this result is readily available in the textbook literature [21]. Alternatively, it (or parts of it) may be assigned as an exercise.

Part b) of this theorem illustrates an initmate relation between linear time-invariant dynamical systems and sinusoidal inputs: the steady state solution of the equation modeling such a system is simply the product of the transfer function evaluated at $j\omega$ and the input. In most input-output analyses the steady state solution is of primary importance since the transient solutions die out when the system is stable.

Equation (3) is important in that it presents us with the key to the development of an experimental technique for the identification of the dynamics of a linear, time-invariant system. Consider the input-output diagram in Figure 7.

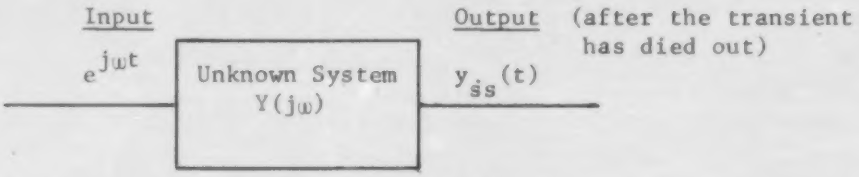


Figure 7. Frequency Response Identification of the Unknown System Y.

In this figure we see that by proper choice of the class of input signals (always a critical decision in a modeling problem), we can construct the complex function $Y(j\omega)$ from Equation (3). Then, it is simply a matter of curve-fitting or parameter estimation to determine the transfer function $Y(s)$ and consequently the dynamics of the unknown system.

3.3 Frequency response (Bode) plots.

Experiments involving frequency response methods for the modeling of an unknown system usually result in frequency response or Bode plots. These are plots of the magnitude and argument (phase) of the complex function $Y(j\omega)$ versus frequency (ω). More precisely, with $M(\omega) = |Y(j\omega)|$ and $N(\omega) = \arg(Y(j\omega))$ (so $Y(j\omega) = M(\omega)e^{jN(\omega)}$) the magnitude plot is the graph of $M(\omega)$ versus ω. It is common engineering practice to use a logarithmic scale $\log_{10}M(\omega)$ versus $\log_{10}\omega$ for the magnitude plot, and phase angle in degrees versus $\log_{10}\omega$ for the phase plot. $\text{Log}_{10}\omega$ is used in place of ω since this allows a much greater range of values to be included, and $\log_{10}M(\omega)$ is used since it makes graphical manipulations somewhat easier (the plot for

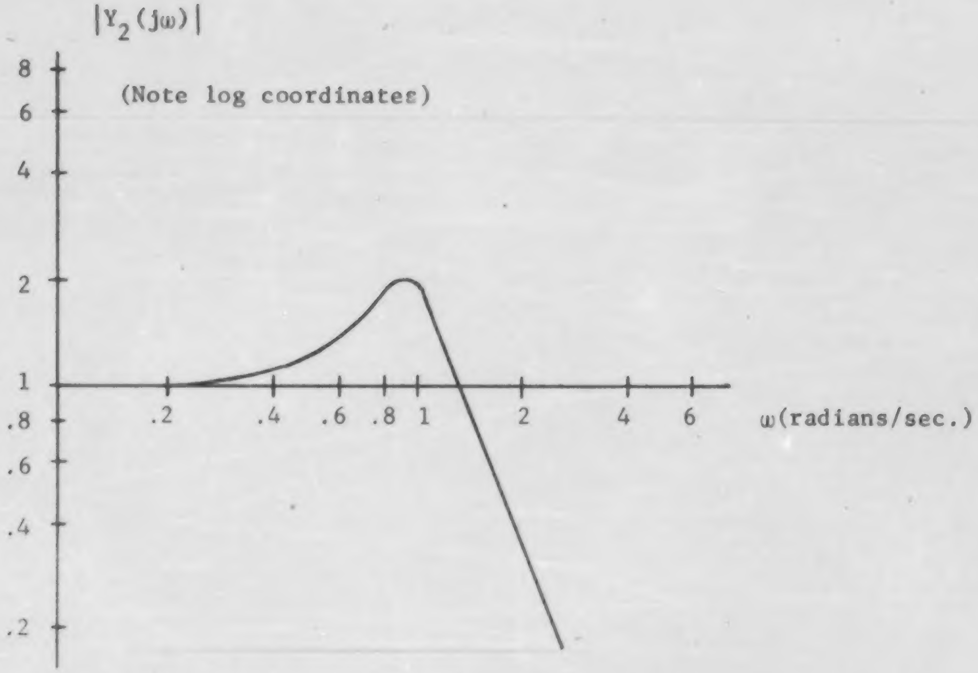


Figure 8. Magnitude Plot of A Second-Order System (Equation 4 with a = 0.25)

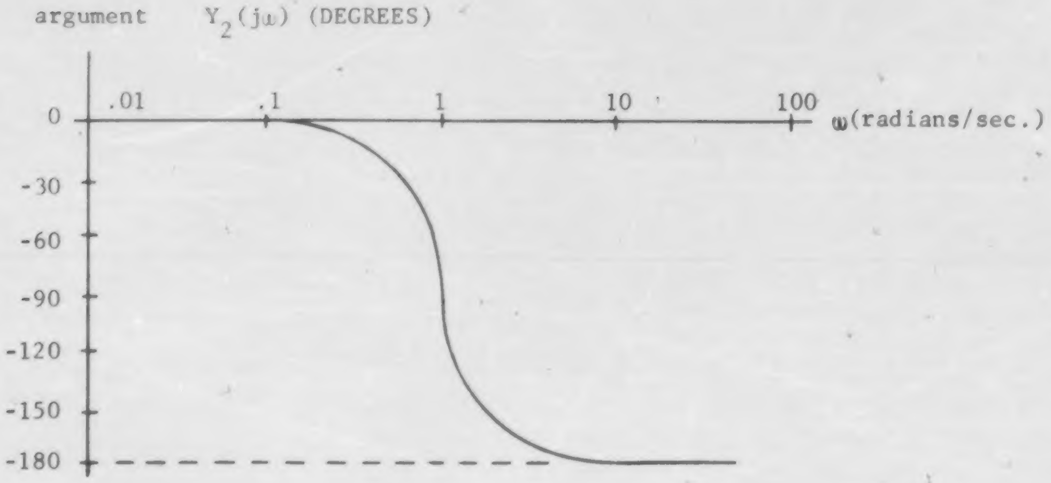


Figure 9. Phase Plot of A Second-Order System (Equation 4 with a = 0.25)

a product of functions becomes the sum of the plots). Amplitude and phase plots for a typical second order transfer function

$$Y_2(s) = \frac{1}{1 + 2as + s^2} \qquad (0 < a < 1) \tag{4}$$

are shown in Figures 8 and 9. (See Chapter 15 of reference 9 for further examples.) After stating the following corollary to Theorem 1(b), we will be ready to attack the rubber band-weight modeling problem.

Corollary: Consider an equation of the form (1) with real coefficients and with $u(t) = \sin \omega t$. Then $y_{ss}(t) = M(\omega)\sin(\omega t + N(\omega))$, where $Y(j\omega) = M(\omega)e^{jN(\omega)}$.

Proof. The steady state solution $y_{ss}(t)$ of $y^{(n)} + a_{n-1} y^{n-1} + \ldots + a_0 y = \sin \omega t$ is the imaginary part of the steady state solution $\tilde{y}_{ss}(t)$ of $y^{(n)} + a_{n-1} y^{(n-1)} + \ldots + a_0 y = e^{j\omega t}$. Thus

$$\tilde{y}_{ss}(t) = Y(j\omega)e^{j\omega t} = M(\omega)e^{jN(\omega)}e^{j\omega t} = M(\omega)e^{j(\omega t + N(\omega))}$$

$$= M(\omega)\left[\cos(\omega t + N(\omega)) + j \sin(\omega t + N(\omega))\right]$$

and $$y_{ss}(t) = M(\omega)\sin(\omega t + N(\omega)).$$

The meaning of this corollary is that if the input to the system is $\sin \omega t$, then the output has frequency ω, amplitude $M(\omega)$ and phase shift $N(\omega)$.

3.4 Linearity and Time Invariance.

The mathematical results noted in § § 3.2 and 3.3 are valid only for linear, time-invariant differential equations. Since these results lead directly to a procedure for determining a model for the dynamics of such an unknown system from input and output data, it would be convenient to have tests to check to see if these conditions (linearity and time-invariance) are satisfied or approximately satisfied for certain regions of operation.

In many cases the time-invariance question can be settled a priori by consideration of the system and the environment in which it operates. For example, in the rubber-band-weight experiment none of the parameters of the system (e.g., weight, elasticity of the rubber band) vary with time. However, environmental parameters (e.g., wind) could vary with time. Naturally, one should attempt to perform this experiment in a uniform environment to eliminate the possibility of time variable dynamics.

The question of linearity is more difficult. It is very rare for a physical system to be absolutely linear over a wide range of operating conditions. However, good results can be obtained with linear models if the physical system is near-linear over the range of operating conditions of interest (i.e., the system need not be near-linear for all parameter values). Systems describable by an equation (1) have the property that for each input-output pair (u_1, y_1) and (u_2, y_2)

i) $u_1 + u_2$ is a legitimate input and the output corresponding to this input is $y_1 + y_2$.

ii) for any constant α the input αu_1 is legitimate and the corresponding output is αy_1.

To test whether a system may be described by an equation (1) one can apply various inputs and sums of inputs and observe the outputs. If i) or ii) are true over a certain frequency range of inputs, one may adopt as a working hypothesis that the system may be adequately so described. The corollary to Theorem 1 provides another test. The output to a sinusoidal input of frequency ω must also have frequency ω, although there is generally a phase shift.

3.5 The experimental procedure.

The corollary of §3.3 was formulated so that the mathematical results derived there could be applied directly to an especially convenient experimental input, $\sin \omega t$. To perform the experiment attach one end of the rubber band string to your finger and have an aide to record observations.

1. Move your hand up and down very slowly imagining that your hand is tracing a sine function of constant, very low frequency, e.g., one cycle in four or five seconds. The amplitude of the oscillation should be roughly three to nine inches and as constant as you can make it. The result should be that the weight moves just as your hand does, i.e., the amplitude is the same and the phase change is nearly imperceptible. Plot these values on the frequency response plots i.e., amplitude ratio $\equiv M(\omega) \approx 1$, phase $\equiv N(\omega) \approx 0^\circ$ (where ω very small). The plotted points should correspond roughly to points ① on Figures 10 and 11.

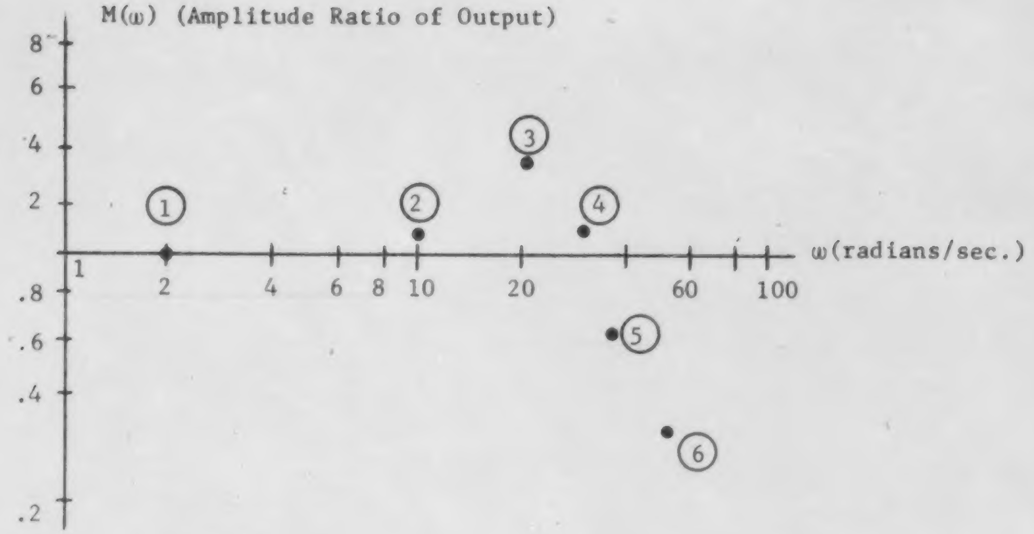


Figure 10. Amplitude Ratio of the Position of the Weight for Sin ωt Input by the Hand

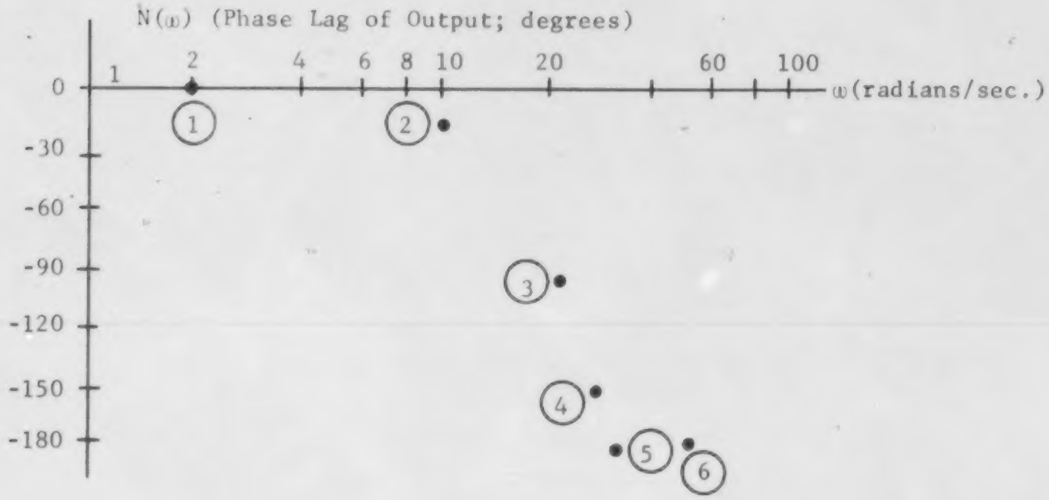


Figure 11. Phase Lag of the Position of the Weight with Respect to the Position of the Hand

2. Gradually increase the frequency of your hand motion but keep the amplitude roughly constant. You will begin to detect a change in the relative motion of the weight with respect to the hand. Estimate roughly this frequency and the resultant amplitude ratio and phase lag. (The amplitude of the weight should be larger than the amplitude of the hand, and the weight should be lagging behind the hand motion.) Plot these points on the frequency response plots; they should correspond roughly to points ② on Figures 10 and 11.

3. Again increase the frequency (with constant amplitude of the hand) in discrete steps, and roughly estimate the frequency, amplitude ratio, and phase lag at the point where the weight has its largest amplitude; the motion of the weight at this point may be somewhat wild. (This is indicating the region of resonance of the system, i.e., the input frequency is close to the natural frequency of the system.) Plot these points on the frequency response plots; they should correspond roughly to points ③ on Figures 10 and 11. If the experiment is done carefully, the phase shift should be almost exactly -90°.

4. Continue as above and plot points for higher and higher frequencies. One can clearly see that the amplitude of the oscillations decreases. However, it is difficult to see (with the eye) that the phase lag is increasing (because of the low-amplitude oscillation of the weight). The phase shift should approach -180° as ω increases (points ④ ⑤ ⑥ on Figures 10 and 11).

Now that the requency response plots are completed, the mathematical model of the system can be constructed. The following theorem is especially helpful in this connection.

Theorem 2. Suppose (1) is stable and that Y is the transfer function. Then

i) $\lim\limits_{\omega\to\infty} M(\omega) = \lim\limits_{\omega\to\infty} |Y(j\omega)| = 0$

ii) $\lim\limits_{\omega\to\infty} N(\omega) = \lim\limits_{\omega\to\infty} (\text{argument } Y(j\omega)) = -\frac{n\pi}{2}$ (radians)

where n is the degree of equation (1).

Proof. Since

$$\lim_{\omega\to\infty} \omega^n Y(j\omega) = j^{-n}$$

it follows that

$$|Y(j\omega)| \text{ is asymptotic to } |\omega^{-n}|$$

from which we conclude that i) holds, and also that

$$\arg(Y(j\omega)) \text{ is asymptotic to } \arg(j^{-n}) = n(-\pi/2)$$

from which we conclude ii).

This theorem gives us a mathematical basis for estimating the order of the differential equation which models our dynamical system provided the high frequency data are fairly accurate (recall the difficulty we had in measuring the high frequency phase angle).

In our example, we estimate the order by determining the asymptotic value of the phase angle in Figure 11. This value is $-\pi = 2(-\frac{\pi}{2})$ which implies that the dynamical system can be approximately modeled by a second order differential equation.

Note: It may be shown (as an exercise) that the slope of the magnitude plot as $\omega \to \infty$ also gives the order of the differential equation:

$$\lim_{\omega\to\infty} \frac{d \log_{10} |Y(j\omega)|}{d \log_{10} \omega} = -n$$

Now that we know a second order differential equation provides an approximate model for the system, we can complete the model by using a least squares curve fit of the magnitude plot and a parameter optimization routine. That is, we have deduced that the equation

$$\frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = u$$

is a good candidate for a mathematical model of the system and we need only determine the parameter values a_0, a_1 which give a good curve-fit of the data in Figure 10.

Note: Bode [10], [12] proved that for stable differential equations of the form (1), the phase plot is uniquely defined by the magnitude plot. Thus, the least squares curve-fit need only be done on the magnitude plot.

3.6 Determination of a_0, a_1 (Least squares and function minimization)

At this point we know that the transfer function has the form $Y(s) = \dfrac{1}{s^2 + a_1 s + a_0}$ for some constants a_0 and a_1, and we have a set of m experimental measurements of the form (ω_i, M_i) where M_i is the observed amplitude for input ω_i. The problem now is to determine the constants a_0 and a_1 in such a way that $M(\omega) = |Y(j\omega)|$ fits the observed data in the best possible way, where "best" remains to be defined.

For each i we have an observed value M_i and a theoretical value $|Y(j\omega_i)|$. The difference $|Y(j\omega_i)| - M_i$ is the "error" in the i^{th} measurement. We want to choose a_0 and a_1 in such a way that the sum of the squares of the errors is minimized. That is, we seek to minimize the function

$$f(a_0,a_1) = \sum_{i=1}^{m} \Big(|Y(j\omega_i;\ a_0,a_1)| - M_i\Big)^2$$

as a function of a_0 and a_1. The choice of least square error as a measure of fit is a common one which can be supported on several grounds. There is also an element of arbitrary choice in selecting this over alternative measures.

If one suspects that the high frequency data is not as reliable as the rest, one might wish to de-emphasize that data by minimizing instead the function

$$(12) \qquad f(a_0,a_1) = \sum_{i=1}^{m} W_i\Big(|Y(j\omega_i;\ a_0,a_1)| - M_i\Big)^2$$

where $W_1, W_2, \ldots, W_n$ are constant weighting parameters. The choice of weights will depend upon the particular data and experimental procedure.

There are many canned computer subroutines available for the determination of a_0 and a_1, given equation (12). Examples are FMCG, FMFP and NEWT in the IBM Scientific Subroutine manual [13]. Alternatively, it is not difficult to implement algorithms based on either the gradient method or Newton's method.

The gradient method begins with a guess $\begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \end{bmatrix}$ and constructs a sequence $\begin{bmatrix} a_0^{(n)} \\ a_1^{(n)} \end{bmatrix}$ based on the formula

$$\begin{bmatrix} a_0^{(J+1)} \\ a_1^{(J+1)} \end{bmatrix} = \begin{bmatrix} a_0^{(J)} \\ a_1^{(J)} \end{bmatrix} - \alpha_J \begin{bmatrix} \frac{\partial f}{\partial a_0}(a_0^{(J)}, a_1^{(J)}) \\ \frac{\partial f}{\partial a_1}(a_0^{(J)}, a_1^{(J)}) \end{bmatrix}$$

The positive scalar α_J is chosen at each step so as to lead to the greatest decrease in f possible at the J^{th} step. See [14], [15] for details of both the theory behind the algorithms and methods for determining α_J.

Newton's method is based on the iteration

$$\begin{bmatrix} a_0^{(J+1)} \\ a_1^{(J+1)} \end{bmatrix} = \begin{bmatrix} a_0^{(J)} \\ a_1^{(J)} \end{bmatrix} - D_J^{-1} \begin{bmatrix} \frac{\partial f}{\partial a_0}(a_0^{(J)}, a_1^{(J)}) \\ \frac{\partial f}{\partial a_1}(a_0^{(J)}, a_1^{(J)}) \end{bmatrix}$$

where $D_J = \begin{bmatrix} \frac{\partial^2 f}{\partial a_0^2}(a_0^{(J)}, a_1^{(J)}) & \frac{\partial^2 f}{\partial a_0 \partial a_1}(a_0^{(J)}, a_1^{(J)}) \\ \frac{\partial^2 f}{\partial a_1 \partial a_0}(a_0^{(J)}, a_1^{(J)}) & \frac{\partial^2 f}{\partial a_1^2}(a_0^{(J)}, a_1^{(J)}) \end{bmatrix}$

In each case, if all goes well, the sequence $\begin{bmatrix} a_0^{(J)} \\ a_1^{(J)} \end{bmatrix}$ will converge to the values $\begin{bmatrix} a_0^* \\ a_1^* \end{bmatrix}$ that minimize $f(a_0, a_1)$. There are advantages and disadvantages to each method:

i) If a poor initial estimate of the parameters is used in a Newton iteration, the sequence may not converge. If the estimate is good, convergence is usually very rapid.

ii) The gradient method will always decrease the function $f(a_0, a_1)$ on successive iterates, but the rate of convergence may be intolerably slow.

The canned algorithms FMCG and FMFP use the conjugate gradient method and the Davidon-Fletcher-Powell formulas, respectively, which were designed to overcome the undesirable properties of the classical gradient and Newton methods. See [14] for the theory behind the algorithms and [15] for some representative simulations.

3.7 Model Verification.

Let us now assume that optimal values $\tilde{a}_0$, $\tilde{a}_1$ of the parameters a_0 and a_1 have been determined. The following questions should be asked:

(1) Is the original assumption that the system is nearly linear a justifiable one? How can it be checked?

(2) What is the physical significance of $\tilde{a}_0$? How can the conjectured significance be checked?

(3) What is the physical significance of $\tilde{a}_1$? How can the conjectured significance be checked?

In model verification, one can only hope to develop necessary conditions, and in this sense, every model is tentative. For example, let us recall how one might answer the first question above. If the system is indeed linear, then the output for the sum of any two inputs $u_1 + u_2$ must be the same as the sum of the respective outputs for the separate inputs u_1 and u_2; and the output for the input αu (α = constant) must be equal to α multiplied by the output for the input u. Since only a finite number of inputs can ever be tested, we only accumulate necessary conditions for the validity of the model with such tests. In practice, one uses such tests to delineate regions of linear operation (e.g., our system is not truly linear, but for a certain frequency range it is very close to linear).

Finally, let us address the second and third questions raised at the beginning of this section, i.e., the physical characterization of a_0, a_1. The system consists of only two different elements: the rubber band string and the weight. Thus, these must be the main contributors to the physical characterization of a_0, a_1. Also, the environment of the experiment must be considered since it will, to some degree, affect the values of a_0, a_1. For example, suppose the same experiment was conducted under water; then the environment of the experiment might not allow us to determine an accurate model. This point is of major importance in the modeling of physiological systems in that the organ to be modeled cannot be isolated in a reasonable experimental environment.

The atmosphere will affect the values of a_0, a_1, but if the weight is heavy enough, this effect can be minimized. Assuming that a_0, a_1 are mainly determined by the rubber bands and weight, let us now determine how they are related. Our first thought might be that each parameter is influenced by only one of the components, e.g., a_0 is influenced only by the weight and a_1 is influenced only by the rubber band's elasticity. To check these hypotheses, the student can vary separately the weight and elasticity of the rubber bands and repeat the estimation of the parameters a_0, a_1. Actually the system is essentially a spring-mass-damper system which is shown (idealized) in Figure 12. The usual model for such a system (based on Newton's laws, see [21] p.2) leads to the equation

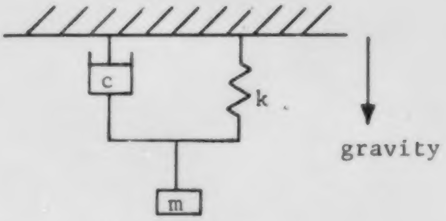


Figure 12. Spring-Mass-Damper Representation of the System.

$$m\,\ddot{y} + c\,\dot{y} + k\,y = 0 \tag{13}$$

or

$$\ddot{y} + \frac{c}{m}\,\dot{y} + \frac{k}{m}\,y = 0, \tag{14}$$

where m = mass of the weight, c = linear damping parameter of the rubber band, and k = linear spring constant of the rubber band. Thus, m is mainly associated with the weight and c, k are associated with the rubber bands. The connection between our parameters a_0 and a_1 and the parameters of this system is

$$a_0 = \frac{k}{m}, \quad a_1 = \frac{c}{m}. \tag{15}$$

Thus, if enough experiments are conducted with various weights and rubber bands of various thicknesses, the data should indicate (roughly) these relationships.

4. PROJECT HUMAN OPERATOR EXPERIMENTAL DATA

In this section data from an actual experiment will be presented for analysis and interpretation. Before presenting the data, one further mathematical property involving a nonlinearity present in all human operators is needed. This nonlinearity is called the inherent time-delay or transport lag, and it is due to the fact that the human operator cannot react instantaneously to a stimulus. (This delay-time is roughly 0.1 to 0.2 seconds for straightforward, compensatory tasks.) Even though this is a nonlinearity, it may be incorporated very easily into linear frequency domain analyses because of the following fact.

Theorem 3. Suppose that the Laplace transform W of u exists for $s \geqq s_0$, and define y by

$$y(t) = \begin{cases} 0, & t \leqq \tau \\ u(t-\tau), & t > \tau. \end{cases}$$

Then the Laplace transform $L[y]$ is related to W by

$$L[y](s) = e^{-\tau s}W(s) \quad \text{for} \quad s > s_o.$$

This is a standard fact concerning Laplace transforms. It means that if y is viewed as the output and u as the input, then y follows u exactly but is delayed by τ units of time.

If a block with a nonlinearity of this type is inserted into a feedback control loop (such as Figure 1), it presents little difficulty if frequency-domain techniques are employed. Indeed, if the transfer function Y is of the form $Y(s) = e^{-\tau s}$, then

$$|Y(j\omega)| = 1, \quad \text{and} \quad \arg Y(j\omega) = -\tau\omega.$$

That is, the magnitude of the frequency response is unaffected by $Y(s)$ whereas the phase is decreased by $\tau\omega$.

4.1 Experimental data

The data for this experiment were developed by J. I. Elkind [16] in one of the earliest comprehensive human operator modeling projects. The basic setup of the experiment is shown in Figure 13. This setup differs from Figure 1 in that no "vehicle" dynamics (or "plant" dynamics) are in the loop.

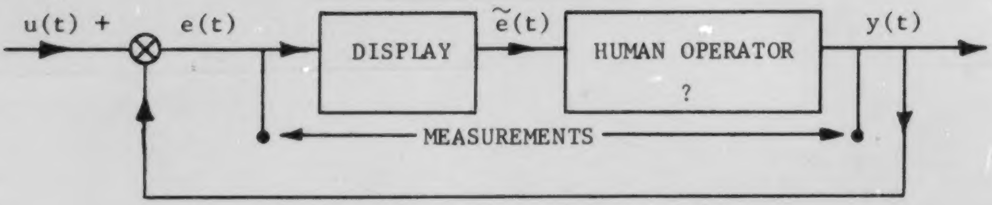


Figure 13. Experiment Setup

Experiments with vehicle dynamics are presented in [2], [3], [8], [17] and [18]. This part of the experiment was mainly concerned with determining transfer function models of the human operator for inputs of various bandwidths, i.e., ranges of frequencies of the components. The magnitude and phase frequency response plots resulting from the experiment are shown in Figure 14.

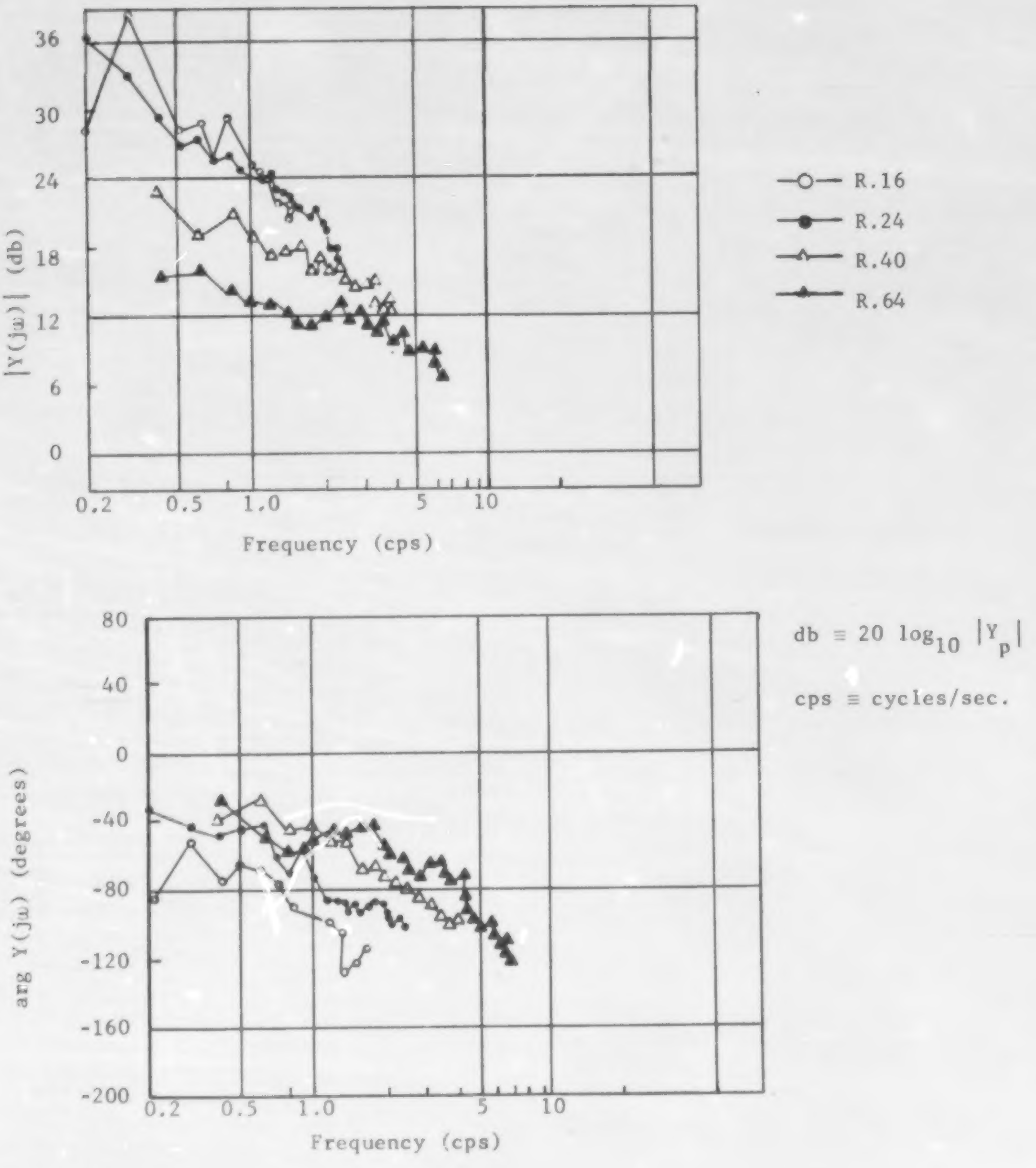
The four curves on each plot correspond to four inputs with different bandwidths, with R.16 and R.64 representing the smallest and largest bandwidths, respectively. The input labelled R.64 has higher frequency components than the other inputs. (The designations, R.16, etc., refer to the fact that the input functions have a rectangular power spectra with the specified cutoff frequency, 16 cps in the case of R.16. See reference 17a for the definitions and details.) The dynamics of the display are neglected so the plots represent the transfer function of the human operator.

Points for discussion and and analysis.

1. Discuss the relation between the human operator's response and the bandwidth of the input. (Rough answer: As the bandwidth increases, the magnitude of the response decreases and the lag in the phase decreases. Thus the operator makes smaller motions if the input has higher frequency components and this probably aids in phase synchronization.

2. What would be a reasonable transfer function for the data of Figure 14? Answer: Elkind proposed

$$Y_{P_1} = \frac{Ke^{-\tau s}}{(Ts+1)} . \tag{16}$$



db ≡ 20 $\log_{10} |Y_p|$

cps ≡ cycles/sec.

Figure 14. Experimentally Derived Frequency Response Plots.
(Figure 29, Reference 17b)

Curve-fitting both the magnitude and phase curves results in the following values [17b]:

Input	τ (SEC)	T (SEC)	K (db)
R.16	0.64	4.55	34.5
R.24	0.264	3.18	31.5
R.40	0.214	1.27	22.5
R.64	0.183	0.58	15.0

(If computer routines are available, a parameter optimization scheme may be employed to solve for the parameters; otherwise rough estimates can be made of the trends of τ, T and K as functions of the input bandwidth.)

These parameter values imply that the "gain" (K) of the human operator decreases with increasing bandwidth, which agrees with intuition. The term $(Ts + 1)^{-1}$ is called a "first-order lag" since it causes an increase in the phase lag, and it has been attributed to a neuromuscular lag, as opposed to the sensory transport lag, τ. Note that T approached 0 as the input bandwidth increased, which implies the human operator tends to behave more as a pure gain and pure transport time-delay with increasing bandwidth.

In [17] it is noted that if the model of equation (16) is inserted into a feedback control loop, the resultant feedback system is unstable at high frequencies. Thus, the following model is proposed to represent a stable feedback system:

$$(17) \qquad Y_{p_2} = \frac{Ke^{-\tau_1 s}}{(T_0 s+1)(T_1 s+1)}$$

With this model, the following values [17a], [17b] are obtained:

Input	τ_1 (SEC)	T_o (SEC)	T_1 (SEC)	K (db)
R.16	.110	4.55	.531	34.5
R.24	.104	3.18	.161	31.5
R.40	.133	1.27	.081	22.5
R.64	.150	0.58	.033	15.0

These values are considered reasonable. Typically, $0.1 \leq \tau_1 \leq 0.2$ and is attributed mainly to the transport lag; T_1 is attributed to neuromuscular lag; and T_0 is a lag which is dependent upon what is being controlled and the bandwidth of the input. Further discussions are presented in [17] and [18].

5. FURTHER TOPICS

The generation of the data presented in Section 4 involved random inputs and correlation methods. The resultant human operator models are more accurately labeled "random input describing functions," the theory of which is presented in [19]. However, they are treated like transfer functions in the analysis of control systems, and in the interpretation of models. A student with knowledge of stochastic processes should be able to follow the presentation in [19] if more insight into the data generation process is desired.

Current research in this area includes among other things: the development of models for more than a single task and the application of optimal control and state estimation theory to the development of models (which assumes the human operator is an optimal information processor and controller subject to inherent constraints). Current results are usually presented at the Annual Conference on Manual Control, which publishes a Proceedings released either by NASA, the United States Air Force, or a sponsoring university (e.g., see NASA SP-144, NASA SP-214).

REFERENCES

1. *Wiener, N. Cybernetics, MIT Press, 1948 and 1961 (paperback)

2. Jex, H. R., et al. "A Study of Fully-Manual and Augmented-Manual Control Systems for the Saturn V Booster Using Analytical Pilot Models," National Aeronautics and Space Administration Report NASA CR-1079, 1968.

3. Weir, D. H. and McRuer, D. T. "A Theory for Driver Steering Control of Motor Vehicles," Highway Research Record, vol. 247, pp. 7-28, 1968.

4. *McRuer, D. T. "Development of Pilot-In-The-Loop Analysis," AIAA Journal of Aircraft, vol. 10, No. 9, pp. 515-522, September, 1973.

5. *Milsum, J. H. Biological Control Systems Analysis, McGraw-Hill Book Company, New York, pp. 242-252 and 418-425, 1966.

6. Bekey, G. A. "Parameter Estimation in Biological Systems: A Survey," in Identification and System Parameter Estimation (P. Eykhoff, editor), American Elsevier Publishing Company, New York, 1973.

7. Eykhoff, P. System Identification, John Wiley and Sons, New York, Chapter 14, 1974.

8. Tustin, A. "The Nature of the Operator's Response in Manual Control and Its Importance for Controller Design," Journal of the Institute of Electrical Engineers (England), vol. 94, Part IIA, No. 2, pp. 190-202, 1947.

9. **DiStefano, J. J.; Stubberud, A. R.; Williams, I. J. Schaum's Outline on Feedback and Control Systems, McGraw-Hill Book Company, New York, 1967.

10. Melsa, J. L. and Schultz, D. G. Linear Control Systems, McGraw-Hill Book Company, New York, 1969.

11. Struble, R. A. Nonlinear Differential Equations, McGraw-Hill Book Company, New York, 1962.

12. Bode, H. W. Network Analysis and Feedback Amplifier Design, Van Nostrand Reinhold Company, Princeton, NJ, 1945.

13. IBM System/360 Scientific Subroutine Package (360A-CM-03X), Version III, IBM Technical Publication Department, 1968.

14. Ortega, J. M. and Rheinboldt, W. C. Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

15. *Himmelbrau, D. M. Applied Nonlinear Programming, McGraw-Hill Book Company, New York, 1972.

16. Elkind, J. I. "Characteristics of Simple Manual Control Systems," Techanical Report No. 111, MIT Lincoln Laboratory, April 6, 1956.

17. *a.) McRuer, D. T. and Krendel, E. S. "The Human Operator as a Servo System Element," Journal of the Franklin Institute, vol. 267, Nos. 5 and 6, pp. 381-403 and pp. 511-536, May and June, 1959.
b.) USAF Wright Air Development Center Report 56-524, October, 1957. (The details for Reference 17a are contained in this report.)

18. *a.) McRuer, D. T.; Graham, D.; Krendel, E. S. "Manual Control of Single-Loop Systems," Journal of the Franklin Institute, vol. 283, Nos. 1 and 2, pp. 1-27 and pp. 145-170, January and February, 1967.
b.) USAF Flight Dynamics Laboratory Report AFFDL-TR-65-15, July, 1965. (The details for Reference 18a are contained in this report.)

19. *Graham, D. and McRuer, D. T. Analysis of Nonlinear Control Systems, Dover Publications, New York, Chapter 6, 1961 (hardback) and 1971 (Dover Paperback).

20. Rabenstein, A. Introduction to Ordinary Differential Equations, Academic Press, New York, 1972.

21. Braver, F. and Nohel, J. A. Ordinary Differential Equations, W. A. Benjamin, New York, 1967.

(Note: ** indicates essential reference; * indicates readily available background reference.)

Chapter 10
NETWORK ANALYSIS OF STEAM GENERATOR FLOW

T. A. Porsching
University of Pittsburgh

TO THE INSTRUCTOR

The contents of this paper should be accessible to a student who has had courses in the elementary calculus and linear algebra. Some familiarity with the basic conservation laws of Physics (i.e., mass, energy and momentum) is desirable, but not essential. The same is true of computer programming, since a student should be extremely gratified by utilizing a computer to administer the coup-de-grace to a problem on which he has spent considerable time.

Section 1: The opening section of the paper attempts to motivate the whole study by relating it to an actual problem encountered in the operation of a nuclear power plant. The instructor should emphasize that this problem is a real concern to the nuclear power industry and involves plant components costing millions of dollars. He should also show the students schematics of the plant and steam generator, and in general terms describe the heat transfer and hydraulic processes which take place. More details of plant operation may be found in [1] and "Systems Summary of a Westinghouse Pressurized Nuclear Power Plant" by G. Masche, which can be obtained by writing to Westinghouse Electric Corporation.

After Section 1 has been discussed, the class will have been presented with a general, non-technical statement of the problem. At this point the instructor may wish to consider other methods of attack in addition to the network approach. For example, if the mathematical sophistication of the class is sufficient to allow consideration of partial differential equations, then the problem can be restated in terms of finding an appropriate solution of the classical equations of fluid dynamics--the momentum (Navier Stokes) equations, and the continuity equation. This is the most straightforward method of attack, but it should be quickly realized that because of the geometry of the flow region and the complex nature of the equations themselves, such an attack is not likely to succeed. As an alternative to a complete solution of the equations, one might settle for a numerical solution by finite differences. This opens the door on a whole new subject--Computational Fluid Dynamics. In this connection the recent Computational Fluid Dynamics, by P. J. Roach, Hermosa Publishers, 1972, may be consulted. However, the direct finite difference approach leads to an enormous computational problem whose solvability is in turn open to question.

Section 2: Section 2 lays the ground work for the development of a model which is based on simple network concepts. This section is autonomous and could also be used to supplement courses in graph theory or the numerical solution of systems of nonlinear equations.

The physical significance of Kirchhoff's node law as a discrete conservation law should be emphasized. It is, of course, the same law that is satisfied by the current in electrical networks.

It would be instructive for the class to construct the incidence matrices of several simple graphs. The proof on page 380 that an incidence matrix with n rows has rank n-1 is not an empty academic exercise. This fact is used later (on page 386) to complete the proof of an existence and uniqueness theorem for network flows. The determination of the rank could be assigned as a class exercise. This is also true of the demonstration on page 380-381.

In electrical network terminology a link characteristic is the familiar Ohm's law. The instructor can motivate the form of equation (2.6) by reminding the class that many flows in physics have been observed to depend on the difference of values of other variables; for example, heat on a temperature difference, electrical current on a voltage difference, and air flow on a difference in atmospheric pressure.

It is extremely important to have the class understand the nature of the network equations (2.7). On page 382 these have been written out in detail for a specific network. The instructor might carry this example further by substituting some simple functions for the link characteristics.

To save time the proof of theorem 2.1 in the case of linear link characteristics (pp. 385-386) may be omitted. However, it does constitute a useful application of matrix algebra.

Section 3: This section develops the network model of the steam generator flow problem. The procedure described on pp. 387 and 388 is essentially the "control volume" approach which is quite common in fluid dynamics.

Some other factors which contribute to the pressure drop in addition to elevation and friction (page 388) are: sudden expansions or contractions of the flow region and sharp bends or corners in the flow region. (See [9] for more details.)

Of the empirical quantities introduced in this section, the friction factor is by far the most elusive. Relative to this, reference [9] may be consulted for more details. Alternately, the instructor could invite an engineering oriented colleague to give a guest lecture on the nature and determination of friction factors and other correlations.

To give the student some appreciation of the complexity of the real problem and the magnitude of the assumptions required to formulate it in terms of a network model, the instructor should discuss the nature of the thermal effects as given on pp. 390-392. On the other hand the appraisal of the link characteristics (pp. 392-394), although it provides some understanding of the pressure drop effects not taken into account, may be omitted if the class is not familiar with partial differential equations.

Section 4: This section describes an iterative method which may be used to solve the network equations numerically. Like section 2, it is autonomous and constitutes a small exercise in numerical analysis. The only notions employed are those of limit, continuity, function composition and the simple bisection method for solving single equations. The bisection method or an equivalent is likely to be a part of any numerical analysis subroutine package supplied by computer manufacturers.

The ideas of convergence rate and successive overrelaxation mentioned on pp. 397 and 398 are not essential for the solution of the network problem. They are intended to provide further avenues of exploration for the curious.

Section 5: Herein is presented a contrived but nontrivial illustrative problem and its numerical solution. Ideally, the class should write a computer program to solve this problem or one that they themselves have developed. If time or computer is not available to allow such an ambitious undertaking, the class should at least hand calculate a few pressure iterates at a given node, assuming that the pressures at the remaining nodes are those of Table 2 (page 401). In this way they will see (hopefully) that the iterates do indeed tend to the appropriate quantity in Table 2.

After the flow problem has been solved, the instructor may wish to refer back to the simple particle deposition model introduced in section 1 and have the class compute some particle deposition fluxes, for a variety of particle concentrations. Furthermore, if it is assumed that the steam generator must be shut down for cleaning when the average deposit thickness exceeds a preselected amount, (say .25 feet), then calculations could be done to determine when this occurs.[1] This is but one way in which the answers of the flow problem can be used to provide relevant information about the actual steam generator.

---

1. For example, one could assume that only particles 1 micron in diameter are involved, that their density is 5 pounds/ft^3 and that their average concentration is 10^{-4} pounds/ft^3. The calculation could then be performed using the data of Table 3 (page 401) and figure 9. Note that the units on the ordinate of this figure are cm/sec and not ft/sec.

1. ORIGIN OF THE PROBLEM

In this first section we want to examine the role of the steam generator in the overall operation of a nuclear power plant. In this way we hope to put into perspective the mathematical problem which will eventually evolve, and at the same time to emphasize its importance.

Figure 1 shows the main flow paths in a nuclear power plant. Using this figure, it is possible to trace the way in which the thermal energy generated by the fissioning of the nuclear fuel is converted into the mechanical energy of the turbine. Notice that there are independent fluid circuits called primary and secondary loops. The fluid in the primary loops is pumped through the reactor core where it undergoes a temperature rise of typically $40\text{-}70^{\circ}$ F. From the core the fluid passes through the hot leg by piping and into the tubes of the steam generator. Here, its acquired heat is transferred through the tube walls to the shell side of the steam generator. The cooler ($\sim 500^{\circ}$ F.) fluid leaving the steam generator is then pumped through the cold legs and returned to the reactor core.

Water in the secondary loop enters the shell side of the steam generator and, passing along the outside of the hot tubes, is converted into steam. The thermal energy of this steam is then expended in the form of mechanical work which drives the turbines. Once through the turbines, the condensed steam is returned to the steam generator.

Figure 2 shows a cross section of a steam generator. To obtain some idea of the data associated with this device we refer to that reported in [1, Ch. 23] for the steam generators of the Oconee Nuclear Station located in South Carolina. These two units are approximately 73 ft. long and 12 ft. in diameter. Each contains about 15,500 tubes and weighs about 600 tons.

Since the tubes are over 60 ft. long and only about 3/4" in diameter, it is obviously necessary to support them at intermediate points along their length. The detail in Figure 2 shows a portion of one of these tube support plates. Notice the three areas where the holes have been broached to provide flow passages for the secondary fluid.

"The principal maintenance problem associated with these heat-exchanger-type steam generators is tube leakage, which can be caused by chemical or mechanical action or a combination of the two" [1, pg. 36-19]. This is an important concern regarding nuclear steam generators because the tubes contain the radioactive reactor coolant. Although the exact mechanism which causes tube pitting is not fully understood, it is strongly correlated with the chemistry and flow distribution of the shell side fluid. For example, there is evidence that periodically, parts of the tubes are not blanketed with the steam-water mixture of the shell side flow. This condition is called "dryout" and is believed to be a key factor in causing tube damage. Also, in regions of low flow, particles may precipitate out of the fluid and form caustic deposits on the tube surfaces. Indeed, it has been argued [19], [20] that particle deposition on a surface is governed by an equation of the form $N = KC$, where N (mass/time × area) is the particle deposition flux, C (mass/volume) is the average particle concentration, and K (length/time) is the deposition coefficient. This last quantity is a function of particle size and local fluid velocity. From Figure 9, which shows a typical plot of K, we see that K depends strongly on this velocity. Utilizing N, it is a simple matter to calculate the buildup of particles of given size on a surface. For if we assume that N is constant, then Nt represents the mass of these particles which settle onto a unit area of surface in t units of time. Hence if the surface is assumed to be clean at $t = 0$, then h, the thickness of the deposit at time t, is $h = Nt/\rho = KCt/\rho$, where ρ is the density of the particulate matter. Since C and ρ are generally available, the ability to determine deposity buildup depends essentially on the ability to determine the fluid velocity.

In view of the above considerations it is important to obtain a realistic approximation of the shell side flow distribution. Due to the complicated 3-dimensional geometry of the shell side, the presence of the two phase steam-water mixture, and the nonlinearities inherent in the conservation laws which govern the flow, it does not appear to be possible to solve this problem in its fullest generality. However, by utilizing certain simplifying assumptions and adopting a particular point of view, we can obtain a tractable mathematical model which, nevertheless, provides an acceptable numerical solution of the flow problem.

In the following sections we shall develop a simplifed version of this model and present a numerical method for its solution. Although the material presented here is intended for classroom instruction, we wish to emphasize that the ideas lie at the heart of a more elaborate method which can be used to analyze flows in steam generators.

2. SOME ELEMENTS OF NETWORK THEORY

As we have previously mentioned, the analysis of shell side steam generator flow is an extremely complex problem. Our approach will be to reduce the continuous problem to a discrete one in the sense that we shall consider flows along certain preselected flow paths. In effect we shall lump the flow in a region of the steam generator into a single average flow for that region and then formulate laws which these lumped flows should obey. The interconnected system of flow paths on which the discrete flows occur is termed a *network*. The literature on networks is already quite large and continues to grow. In this section we shall deal only with those few notions which are necessary for the treatment of our problem. For further reading on the subjects of networks, graphs,[1] and other discrete flow problems see [2]-[5], [21].

We abstractly define our network η as a couple (V,S). Here V is a finite set of unordered elements called *nodes* and S is a set of ordered pairs of elements of V called *links*. In modeling our flow problem the nodes will represent junctions where the flow changes direction and the links will define the flow paths. We assume that there are n elements in V and m elements in S.

Since the nodes are isomorphic to the first n positive integers, we let $V = \{1, \ldots, n\}$. On the other hand the j^{th} link s_j of η is denoted by $(P(j), Q(j))$ where $P(j), Q(j) \in V$. Thus we can write $S = \{s_j = (P(j), Q(j)) | j = 1, \ldots, m\}$. The nodes $P(j)$ and $Q(j)$ are termed respectively the *initial* and *terminal* nodes of link s_j and constitute the *extremities* of the link. At the same time link s_j is said to be *incident* upon nodes $P(j)$ and $Q(j)$.

---

[1] Although the distinction between networks and graphs is fuzzy, we will use the term "network" to emphasize the idea of a system which carries flow. "Graph," on the other hand, suggests a skeleton of connections without any particular flow connotation.

These algebraic definitions are geometrically motivated. By drawing the nodes as numbered circles and the links as arcs on which numbered arrow heads have been placed pointing from the initial to the terminal node, we obtain a complete description of the network. For example, in Figure 3 we have depicted the network defined by the sets

$$V = \{1, 2, 3\},$$

$$S = \{(3,1), (1,2), (3,2), (3,1), (1,1)\}.$$

Note the presence of the self-loop $s_5 = (1,1)$ and the parallel, but distinct, links $s_1 = (3,1)$, $s_4 = (3,1)$. Since self-loops do not generally represent physically meaningful flow paths, we shall exclude them from further consideration.

A sequence of network links $\{s_{i_1}, s_{i_2}, \ldots, s_{i_k}\}$ is a chain if for $j = 2, \ldots, k-1$, link s_{i_j} has one extremity in common with $s_{i_{j-1}}$ and the other with $s_{i_{j+1}}$. Nodes p and q are connected if there is a chain $\{s_{i_1}, \ldots, s_{i_k}\}$ such that they are extremities of links s_{i_1} and s_{i_k} respectively. The network itself is connected if every pair of nodes in it is connected. Since liquid flows in an inherently continuous manner,[1] we shall confine our attention to connected networks.

In modeling our flow problem we shall make use of the fundamental law of conservation of mass. The network analog of this law is known as Kirchhoff's node law and is mathematically formulated as follows. With each node i of $\mathcal{N}$ we associate two sets of links,

$$\omega^+(i) = \{(i,k) \mid k \in V, (i,k) \in S\},$$

and

$$\omega^-(i) = \{(k,i) \mid k \in V, (k,i) \in S\}.$$

Loosely speaking, $\omega^+(i)$ is the set of links incident upon node i which point away from the node, while $\omega^-(i)$ contains those which point toward the node. With each link s_i of $\mathcal{N}$ we associate a real number w_i. Then the m-dimensional vector[1] $w = (w_1, \ldots, w_m)^T$ is a *flow on $\mathcal{N}$* if

---

[1] There is at least one notable exception to this. See [6].

$$\sum_{s_j \in \omega^+(i)} w_j - \sum_{s_j \in \omega^-(i)} w_j = 0, \quad i = 1, \ldots, n. \tag{2.1}$$

Equation (2.1) expresses Kirchhoff's node law which states that the net flow into a node is equal to the net flow out of it. If w is a flow, then w_j is the flow on link s_j. Actually, it would be more precise to call the w_j link flow *rates* since their units turn out to be mass per unit time, e.g., pounds/sec. Notice that the link flows may assume both signs. We adopt the usual convention that $w_j > 0$ means that the *actual* direction of flow on link s_j is from the initial node to its terminal node, while $w_j < 0$ means that the actual direction of flow is from terminal to initial node.

There is a convenient, compact way in which to write (2.1). This involves introducing the (node-link) incidence matrix. This is the $n \times m$ matrix $A = [a_{ij}]$ where

$$a_{ij} = \begin{cases} +1 & \text{if } i = P(j), \\ -1 & \text{if } i = Q(j), \\ 0 & \text{otherwise} \end{cases} .$$

For example,

$$A = \begin{bmatrix} -1 & 1 & 0 & -1 \\ 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \tag{2.2}$$

is the incidence matrix corresponding to the network in Figure 3 when the self-loop s_5 is removed. Utilizing this matrix, it is clear that the system (2.1) becomes simply

$$Aw = 0. \tag{2.3}$$

Since any link is incident upon its initial and terminal nodes, and only those nodes, every column of A contains, except for zeros, exactly one +1 and one -1. Therefore the rows of A sum to the zero vector and consequently its row rank is less than or equal to $n-1$. In fact it is exactly $n-1$. To prove this we discard row n of A. Then we note that the links which are incident upon the node set $V_1 = \{n\}$ determine columns containing exactly one nonzero entry. The rows in which these nonzero entries lie cannot occur in any vanishing nontrivial linear combination of the rows of A. Thus we may discard them from A. But these same rows define a set of nodes, say V_2, and we can repeat the above argument on links which are incident upon V_2 and which have their other extremities in the complement of V_2. Because of the assumed connectedness of η, this process will eventually exhaust the first $n-1$ rows of A showing that they are independent.

---

[1] We shall always use superscript T to denote transpose. Thus w is a column vector.

Since the row and column ranks of a matrix are equal (see, for example, [7, pg. 42]), we immediately have that the dimension of the null space of A is $m - (n-1)$. Hence, the most general flow on η depends on $m - n + 1$ arbitrary parameters. Clearly, we need to develop further conditions to single out a particular flow from among the multitude of solutions of (2.3).

In modeling the flow problem we must include a means of treating boundary conditions. These are the conditions which hold at those places where the fluid enters and leaves the region of interest, for example, the regions adjacent to the tube sheets in Figure 2. In a network, boundary conditions are accommodated by designating certain nodes as boundary nodes and prescribing the appropriate conditions at these nodes. Since boundary nodes correspond to points where the fluid enters and leaves the network, Kirchhoff's node law cannot be expected to hold at them. However, as a consequence of imposing this law at all the other nodes of the network, we can show that the sum of the boundary flows is zero. In other words the total flow into the network is equal to the total flow out of it. To prove this statement let us assume that there are $n - \nu > 0$ boundary nodes and that these are numbered $\nu + 1, \ldots, n$. Then we can partition the incidence matrix A as $\begin{bmatrix} A^o \\ \partial A \end{bmatrix}$, where A^o consists of the first ν rows of A and represents the nonboundary or interior nodes, and ∂A consists of the last $n-\nu$ rows of A and represents the boundary nodes. If we define u^o and ∂u respectively as ν and $n-\nu$ dimensional vectors of ones, then we have that since A is an incidence matrix,

$$({u^o}^T, \partial u^T)\begin{bmatrix} A^o \\ \partial A \end{bmatrix} = {u^o}^T A^o + \partial u^T \partial A = 0.$$

Therefore, if w is a flow on η,

$$u^{o^T}(A^o w) + \partial u^T(\partial A w) = 0. \tag{2.4}$$

On the other hand Kirchhoff's node law applies at the interior nodes so that

$$A^o w = 0. \tag{2.5}$$

Then certainly $u^{o^T}(A^o w) = 0$, and it follows from (2.4) that $\partial u^T(\partial A w) = 0$ which is what we wanted to prove.

The network shown in Figure 4 may be used to model part of the flow region in a steam generator. In this network the boundary nodes are nodes 13 - 18. The modeling procedure by which the physical flow region can be reduced to such a network will be discussed in the next section.

As previously noted, Kirchhoff's node law does not in general define a unique flow on a network.

The additional conditions which are required to do this come in the form of relations between the flows and a new set of variables $\{p_i\}$ associated with the nodes and termed node states. In our flow problem these states can be interpreted as the static pressures which exist at points in the flow region. We will assume that the flow-node state relation which holds for each link s_j is of the form

$$w_j = f_j(p_{P(j)} - p_{Q(j)}). \tag{2.6}$$

Here $f_j(t)$ is a function of the variable t and $\Delta p_j \equiv p_{P(j)} - p_{Q(j)}$ is the state difference or pressure drop across link s_j. Equation (2.6) is called a link characteristic and states that the flow on link j is a known function of only the pressure drop across the link.

We are now ready to consider the basic network flow problem. Suppose that we are given a link characteristic (2.6) for each link of a network η which has $n-\nu$ boundary nodes $\nu+1, \ldots, n$ at which the pressures $p_{\nu+1}, \ldots, p_n$, are known. Then we seek to determine pressures $p_1, \ldots, p_\nu$ and flows $w_1, \ldots, w_m$ which satisfy the characteristics (2.6) for $j = 1, \ldots, m$ and Kirchhoff's node law (2.1) for $i = 1, \ldots, \nu$ (i.e., at the interior nodes of η). We observe that if we simply substitute the equations (2.6) into the first ν equations (2.1), then we have precisely ν equations and ν unknowns. Thus, if there is a unique set of pressures which satisfies this system, the equations (2.6) determine a corresponding unique set of flows and the problem is completely solved. It is clear that the equations to be satisfied by the pressures are

$$(2.7) \qquad \sum_{s_j \in \omega^+(i)} f_j(p_{P(j)} - p_{Q(j)}) - \sum_{s_j \in \omega^-(i)} f_j(p_{P(j)} - p_{Q(j)}) = 0, \qquad i = 1, \ldots, \nu.$$

We term equations (2.7) the network equations.

By utilizing the incidence matrix once again, we can write the network equations in a concise form. Let $p = (p_1, \ldots, p_n)^T$ and $\Delta p = (\Delta p_1, \ldots, \Delta p_m)^T$ denote n and m dimensional vectors of pressures and pressure drops. From the manner in which the incidence matrix $A = [a_{ij}]$ has been defined, it is easy to see that $\Delta p_j = p_{P(j)} - p_{Q(j)} = \sum_{i=1}^{n} a_{ij} p_i$, $j = 1, \ldots, m$. Hence $\Delta p = A^T p$, where the matrix A^T is the transpose of A. Letting $f(\Delta p) = (f_1(\Delta p_1), \ldots, f_m(\Delta p_m))^T$, it follows from (2.5) and (2.6) that

$$A^o w = A^o f(\Delta p) = A^o f(A^T p) = 0.$$

Therefore, another form of (2.7) is

$$(2.8) \qquad A^o f(A^T p) = 0.$$

Let us write the first three network equations for the network of Figure 4. Since[1]

$$\Delta p_1 = p_{13}{}^* - p_1, \quad \Delta p_2 = p_{14}{}^* - p_2, \quad \Delta p_3 = p_{15}{}^* - p_3,$$

$$\Delta p_4 = p_1 - p_4, \quad \Delta p_5 = p_2 - p_5, \quad \Delta p_6 = p_3 - p_6,$$

$$\Delta p_{16} = p_1 - p_2, \quad \Delta p_{17} = p_2 - p_3,$$

and since Kirchhoff's node law for nodes 1, 2, 3 reads respectively

$$w_4 + w_{16} - w_1 = 0$$

$$w_5 + w_{17} - w_2 - w_{16} = 0,$$

$$w_6 - w_3 - w_{17} = 0,$$

---

[1] To emphasize that nodes 13-15 are boundary nodes where the pressures are known, we have "starred" these quantities.

we obtain

$$f_4(p_1 - p_4) + f_{16}(p_1 - p_2) - f_1(p_{13}^* - p_1) = 0,$$

$$f_5(p_2 - p_5) + f_{17}(p_2 - p_3) - f_2(p_{14}^* - p_2) - f_{16}(p_1 - p_2) = 0,$$

$$f_6(p_3 - p_6) - f_3(p_{15}^* - p_3) - f_{17}(p_2 - p_3) = 0.$$

Notice that in the first equation, which corresponds to Kirchhoff's law for node 1, the unknown quantity p_1 appears in each term and is multiplied by +1 or -1 as the corresponding flow is added or subtracted. Similar remarks can be made about the remaining two equations. Indeed, this is a property of the network equations in general. Later in Section 4 we shall exploit this important property in constructing an algorithm to numerically solve the equations.

At this point we can ask two fundamental mathematical questions about the network equations:

(I) Do they have a unique solution (indeed, do they have any solution at all)?

(II) If the answer to (I) is yes, how can we compute this solution?

In the remainder of this section we shall try to answer question (I), postponing further consideration of (II) until Section 4.

First of all, unless we say more about the functions $f_j(t)$, the answer to question (I) is NO! For example, suppose that we examine the network equations corresponding to the simple one link network shown in Figure 5. Let us assume that node 2 is the boundary node and that the pressure there, p_2^*, is zero. Then there is only one equation and it reads

$$f_1(p_1) = 0. \tag{2.9}$$

If we are given that $f_1(t) \equiv 1$, then (2.9) has no solution. On the other hand if we are given $f_1(t) = \sin t$, then (2.9) has an infinite number of solutions.

Fortunately, experience has shown that the functions $f_j(t)$ occurring in practice do not resemble either of the above functions. In fact it is reasonable to assume that as we increase the pressure drop (or driving force) across a given link the flow in that link also increases. This leads us to assume that for each j, $f_j(t)$ is a strictly increasing function of t. This is still not enough to guarantee an affirmative answer to (I) since the functions

$$f_1(t) = \begin{cases} t & \text{if } t < 0 \\ 1 + t & \text{if } t \geqq 0 \end{cases} \tag{2.10}$$

and

$$f_1(t) = 1 + e^t \tag{2.11}$$

are both strictly increasing, but neither allows a solution of (2.9). The difficulty with (2.10) is the discontinuity at $t = 0$, while the problem with (2.11) is that its range is restricted to positive numbers. We can remove both of these difficulties by assuming that $f_j(t)$ is a continuous function of t and has an unrestricted range.[1] This leads us to define an admissible characteristic as one for which $f_j(t)$ is defined on $-\infty < t < \infty$, is continuous and strictly increasing there, and satisfies $\lim_{t\to\pm\infty} f_j(t) = \pm\infty$. Note that if $f_1(t)$ defines an admissible characteristic, then (2.9) has exactly one solution. In fact we have the following theorem.

Theorem 2.1. If the links of a network $\mathcal{N}$ have admissible characteristics, and if $\mathcal{N}$ has at least one boundary node, then the network equations (2.7) or (2.8) have a unique solution.

We shall not prove this theorem in its full generality since such a proof involves notions which we are not prepared to introduce. A proof can be found, for example, in [8]. We will prove it, however, in the special case that the functions $f_j(t)$ are linear, i.e., of the form

$$f_j(t) = d_j t + c_j, \tag{2.12}$$

where d_j and c_j are constants. These define admissible characteristics if and only if $d_j > 0$.

Let $D = \text{diag}(d_1, \ldots, d_m)$, i.e., the diagonal matrix whose i^{th} diagonal elements is d_i. Also let $c = (c_1, \ldots, c_m)^T$ and assume that the last $n - \nu > 0$ nodes are the boundary nodes. If we go back to the form of the network equations given by (2.8), we see that the i^{th} equation is

---

[1] It is quite natural to assume that the flow in any link is a continuous function of the pressure drop across that link. However, the assumption that the flow becomes infinite with its pressure drop is somewhat artificial since it is more likely that the link will saturate at a finite flow.

$$\sum_{j=1}^{m} a_{ij} d_j (A^T p)_j + \sum_{j=1}^{m} a_{ij} c_j = 0, \quad i = 1, \ldots, .$$

or

(2.13) $$A^o D (A^T p) + A^o c = 0.$$

By writing

$$p^o = (p_1, \ldots, p_\nu)^T,$$

$$\partial p = (p_{\nu+1}{}^*, \ldots, p_m{}^*)^T,$$

$$p = \begin{bmatrix} p^o \\ \partial p \end{bmatrix},$$

$$A = \begin{bmatrix} A^o \\ \partial A \end{bmatrix},$$

we can split off the boundary dependent part of (2.13). In fact we see that (2.13) may now be written as

$$A^o D [A^{o^T} \vdots (\partial A)^T] \begin{bmatrix} p^o \\ \partial p \end{bmatrix} + A^o c = 0,$$

or

$$A^o D [A^{o^T} p^o + (\partial A)^T \partial p] + A^o c = 0,$$

or finally

(2.14) $$(A^o D A^{o^T}) p^o = -(A^o c + A^o D (\partial A)^T \partial p).$$

Equation (2.14) represents a nonhomogeneous system of ν linear equations in the ν unknown pressures of the vector p^o (note that the right hand side of (2.14) is known). By the fundamental theorem on the solvability of linear equations, (2.14) has a unique solution if and only if the $\nu \times \nu$ coefficient matrix $A^o D A^{o^T}$ is nonsingular.

We show that $A^o D A^{o^T}$ is nonsingular by assuming the contrary. Then there is a vector $x = (x_1, \ldots, x_\nu)^T \neq 0$ such that $A^o D A^{o^T} x = 0$. Thus $x^T A^o D A^{o^T} x = 0$. If we let $z = (z_1, \ldots, z_m)^T = A^{o^T} x$, this last equation becomes

(2.15) $$z^T D z = 0.$$

Since D is a diagonal matrix, we can write (2.15) in longhand as

$$\sum_{j=1}^{m} d_j z_j^2 = 0. \tag{2.16}$$

But each $d_j > 0$, so that the only way in which (2.16) can hold is for $z_j = 0$, $j = 1, \ldots, m$. Hence $z = 0$. Now according to the definition of z, it is a linear combination of the columns of A^{o^T}, i.e., the rows of A^o. In fact if we denote the i^{th} row of A^o by a_i then we have

$$z = \sum_{j=1}^{\nu} a_i x_i = 0. \tag{2.17}$$

But we have seen that the rank of A is n-1. Therefore, the vectors a_i, $i = 1, \ldots, \nu \leqq n-1$ are linearly independent and (2.17) implies that $x_i = 0$, $i = i, \ldots, \nu$, which is a contradiction. This establishes the nonsingularity of $A^o D A^{o^T}$ and completes the proof.

We remark once more that although the above proof applies only to the linear case (2.12), the theorem is true for a much wider class of functions. For example, if $b_j > 0$, $d_j > 0$ and c_j are constants, then

$$f_j(t) = b_j t^3 + d_j t + c_j.$$

defines an admissible nonlinear characteristic. The characteristics we shall introduce in the next section to model the steam generator flow problem will be nonlinear.

3. MODELING THE FLOW PROBLEM

We are now ready to show how to model the steam generator flow problem so that the network equations of Section 2 apply. Although the development will be quite general, it will be helpful to apply the ideas to a specific example. Suppose then that we wish to determine the flow distribution in a portion of the shell side of the steam generator pictured in Figure 2. This region, which is shown in Figure 6, page 405, is symmetric about one of the tube support plates and extends to the adjacent plates. For simplicity, we have not shown the tubes, but one must remember that they occupy the region and are vertically oriented. Furthermore, although the actual flow region is three dimensional, the region in Figure 6 is a two dimensional "slice" taken at the symmetry plane of the steam generator as shown in the plan view of the figure.

Now let us divide this region into a finite number of subregions or cells by inserting fictitious boundaries. We have elected to use twelve such cells, and these are numbered by associating nodes with them in the manner of Figure 6. We have indicated the fictitious boundaries by dashed lines.

Since movement of fluid between two adjacent cells must occur across their common boundary, we can account for the flow communication between them by lumping the distributed flow across the whole boundary into a single quantity which represents the total flow across the boundary. It is natural to geometrically signify this flow communication by inserting a link between adjacent cells. Thus in Figure 6 we connect the nodes of the twelve cells by inserting seventeen links. To recognize that the fluid enters and leaves the region across the horizontal boundaries of cells 1, 2, 3, and 10, 11, 12, we connect each of these nodes to a boundary node. The geometric realization of this procedure is the network of Figure 4.

Consider cell i. According to the principle of conservation of mass, if there is no fluid being created or destroyed in the cell and if the density of the fluid in the cell is constant, then the total mass efflux across the boundaries of the cell is equal to the total mass influx. But the movement across boundaries can now be thought of in terms of flows on the links of an associated network. Thus if we let w_j denote the mass flow, say in pounds per second, across the boundary which is penetrated by link s_j, and if $w_j > 0$ corresponds to fluid movement from the initial to the terminal node of s_j, then for cell i we have:

$$\text{Total mass efflux} = \sum_{\substack{s_j \in \iota^+(i) \\ w_j > 0}} w_j - \sum_{\substack{s_j \in \iota^-(i) \\ w_j < 0}} w_j$$

$$\text{Total mass influx} = \sum_{\substack{s_j \in \iota^-(i) \\ w_j > 0}} w_j - \sum_{\substack{s_j \in \iota^+(i) \\ w_j < 0}} w_j$$

Therefore,

$$\sum_{\substack{s_j \in \omega^+(i) \\ w_j > 0}} w_j + \sum_{\substack{s_j \in \omega^+(i) \\ w_j < 0}} w_j = \sum_{\substack{s_j \in \omega^-(i) \\ w_j > 0}} w_j + \sum_{\substack{s_j \in \omega^-(i) \\ w_j < 0}} w_j$$

or

$$\sum_{s_j \in \omega^+(i)} w_j - \sum_{s_j \in \omega^-(i)} w_j = 0,$$

which is Kirchhoff's node law for node i.

The static pressure will in general assume different values at the different nodes of the cells. The pressure difference between two nodes of adjacent cells--which may be regarded as the pressure drop across the link connecting the two cells--is due to a number of factors. We shall consider only two of these:

(I) The effect of the fluid's weight.

(II) The frictional resistence of the tube surfaces and other impervious boundaries on the moving fluid.

If we refer to Figure 7, page 406 , then the weight of the fluid in link s_j causes the pressure at node $P(j)$ to exceed that at node $Q(j)$ by an amount given by

$$(\Delta p_j)_{el} = \rho L_j g \cos \theta_j . \tag{3.2}$$

Here ρ is the (mass) density of the fluid, g is the acceleration due to gravity, L_j is the length of the link and θ_j is the angle at which the link is inclined to the vertical. Notice that we have used the subscript "el" to denote this pressure drop since it is due to the difference in elevation of the link's extremities.

As the fluid moves over the links, it passes around the tubes, along the shroud and through the broached openings in the tube support plates. Each of these exerts a frictionally resistive force on the fluid which is balanced by a pressure difference across the link's extremities. Extensive experimentation has empirically established the form of this pressure drop as (see [9], chapter 11, for more details)

$$(3.3) \qquad (\Delta p_j)_{fr} = \frac{F_j}{\rho A_j^2} |w_j| w_j ,$$

where A_j is the cross sectional area[1] of the link and F_j is a proportionality constant called the friction factor.[2]

Note that if we adopt a constant set of units, for example, if we measure mass in pounds (lb), force in poundals (pd), time in seconds (sec), and length in feet (ft), then the units of the right hand side of (3.2) are

$$\frac{lb}{ft^3}\, ft\, \frac{ft}{sec^2} = [lb\, \frac{ft}{sec^2}]\, \frac{1}{ft^2} = \frac{pd}{ft^2}$$

and those of (3.3) are (F_j being dimensionless)

$$\frac{ft^3}{lb}\, \frac{1}{ft^4}\, \frac{lb^2}{sec^2} = [lb\, \frac{ft}{sec^2}]\, \frac{1}{ft^2} = \frac{pd}{ft^2}$$

Thus in both cases we obtain the units of pressure drop as required.

To obtain the total pressure drop across link s_j we simply add the contributions due to friction and elevation. That is

$$(3.4) \qquad \Delta p_j = (\Delta p_j)_{fr} + (\Delta p_j)_{el} = \frac{F_j}{\rho A_j^2} |w_j| w_j + \rho L_j g \cos \theta_j .$$

Since the friction factor and the density are positive quantities, this equation defines the pressure drop as a strictly increasing, nonlinear function of the link flow. A graph of this function is shown in Figure 8, page 406 . This figure also shows the inverse function obtained by solving (3.4) for w_j. It is easy to see that this inverse is given explicitly by,

---

[1] Since the links are conceptual devices it is somewhat ambiguous to speak of their cross sectional areas (the same is true of their lengths, but it is natural to take these to be the distances between the nodes). One way to assign cross sectional areas is to divide the volumes of the cells by the number of links incident upon them, assign these subvolumes to the appropriate links and then define the cross sectional area to be the ratio of a link's volume to its length. For example, if s_{16} connects nodes 1 and 2 in Figure 6, then $A_{16} = \frac{1}{L_{16}}(\frac{\text{volume of cell 1}}{3} + \frac{\text{volume of cell 2}}{4})$.

[2] The definition of friction factor varies from author to author. Our definition has been motivated by a desire to produce a simple, but realistic, form of the frictional pressure drop $(\Delta p_j)_{fr}$.

$$w_j = f_j(\Delta p_j) \equiv A_j \sqrt{\frac{\rho}{F_j} \left| \Delta p_j - \rho L_j g \cos \theta_j \right|}\ \operatorname{sgn}(\Delta p_j - \rho L_j g \cos \theta_j), \tag{3.5}$$

where sgn(x) is the so called sign function and is defined by

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0\,. \\ -1 & \text{if} \quad x < 0 \end{cases}$$

Now we observe that (3.5) is precisely of the form (2.6) and so defines a link characteristic. Moreover since (3.1) is Kirchhoff's node law, we see that we have reduced the flow problem to that of solving the network equations. Not only does (3.5) define a link characteristic but what is even better, it defines an admissible characteristic. Therefore, by Theorem 2.1, if we have at least one boundary node in our network, then the flow problem has a unique solution.

We have arrived at this fortunate mathematical state of affairs by simplifying certain aspects of the actual problem. One of the most crucial simplifying assumptions we have tacitly made is to neglect the influence of heat. Recall that in the early discussion of the steam generator, we mentioned that the flowing quantity of interest is a two phase mixture of vapor and water. Now the modeling of a two phase flow process is in general quite complicated and utilizes some assumptions which cannot be derived from first principles (that is, they are empirical correlations which have evolved through experimentation). Possibly the simplest approach to the problem is to assume that w_j represents the mass flow rate of the two phase mixture, and that the fraction of this which is vapor--this fraction is term the quality--is a simple function of the thermal energy of the mixture. Among other things this assumption implies that the two phases move together at the same velocity, a condition which is known to be violated for certain types of slow motion.

One measure of the thermal energy of the mixture is the enthalpy H. Under the above assumption, if we know the mass flow and enthalpy of the mixture, then we know (via the quality) the mass flow of the vapor present in the mixture. The introduction of the new variable H requires the addition of another equation to the model. This comes from the principle of Conservation of Thermal Energy and for the network model we have been considering it may be written

(3.6)

$$H_i\Big(\sum_{\substack{s_j \in \omega^+(i) \\ w_j > 0}} w_j - \sum_{\substack{s_j \in \omega^-(i) \\ w_j < 0}} w_j \Big) = \sum_{\substack{s_j \in \omega^-(i) \\ w_j > 0}} \Big(w_j H_{P(j)} + \varphi_j \Big) - \sum_{\substack{s_j \in \omega^+(i) \\ w_j < 0}} \Big(w_j H_{Q(j)} - \varphi_j \Big).$$

$i = 1, \ldots, \nu.$

Here we have denoted the external rate of heat addition[1] to link j by φ_j. Note that like the pressure, the enthalpy is a node variable.

If we measure heat in British Thermal Units (BTU), then the units of enthalpy can be taken as BTU/lb. If the units of φ_j are BTU/sec, then both sides of (3.6) have units of BTU/sec. Thus (3.6) equates heat rates at node i. The left side is the rate at which heat is being removed from the node by fluid efflux and the right side is the rate of heat addition to the node by the incoming fluid.

We notice the presence of the flow rates w_j in (3.6). If we presume that we know these, and if we are given the φ_j (that is, if we know the heat addition from the tubes), then equations (3.6) constitute a set of linear equations for the enthalpies. On the other hand because the enthalpies do not appear in the network equations, we can solve them without reference to equations (3.6). This is a consequence of our assumption that the density ρ is constant. In the more general case, the density at any point in the flow region is a function of the pressure and enthalpy at the point, say

(3.7) $$\rho = R(p, H).$$

Equation (3.7) is known as the equation of state. In steam generator calculations, the pressure variation on the shell side is sufficiently small so as to justify regarding ρ as being independent of p. In other words ρ may be evaluated at a constant "system pressure" p^* so that (3.7) becomes[2]

(3.8) $$\rho = R(p^*, H).$$

---

[1] In our steam generator model this is, of course, the heat which is conducted through the tube walls from the primary loop.

[2] See footnote on pg. 401.

If H were constant, then indeed the density would also be constant. However, the assumption of constant H is difficult to justify, especially with the formation of vapor. If we assume that ρ is given by (3.8) and introduce this into the characteristic (3.5), then the network equations contain the node enthalpies and must be solved simultaneously with equations (3.6). This is a much more difficult problem than the one involving only the network equations and goes well beyond the instructional intent of this paper. Therefore, we shall avoid it by assuming that the density is constant.

We shall devote the remainder of this section to an appraisal of link characteristic (3.5) by examining its equivalent form (3.4). This requires the introduction of certain partial differential equations but does not affect either the network equations or the subsequent development. Therefore, the continuity of the presentation will not be interrupted if the rest of the section is omitted.

The application of Newton's second law to an infinitesimal element of fluid results in a set of partial differential equations known as the Navier-Stokes equations. A derivation of these equations can be found in almost any book on hydrodynamics, for example [9] or [10]. If the motion is steady[1] and two dimensional, as in the case of the steam generator slice considered earlier, then these equations are

(3.9)
$$\rho\left(u\,\frac{\partial u}{\partial x} + v\,\frac{\partial u}{\partial y}\right) = -\,\frac{\partial p}{\partial x} + \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + F_1,$$

(3.10
$$\rho\left(u\,\frac{\partial v}{\partial x} + v\,\frac{\partial v}{\partial y}\right) = -\,\frac{\partial p}{\partial y} + \mu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) + F_2.$$

Equations (3.9) and (3.10) are written in terms of the rectangular coordinates (x,y). The quantities u and v denote the fluid velocity components in the x and y directions respectively, p is the pressure, and ρ is the (constant) density. The constant μ is called the viscosity and reflects the fact that elements of a moving fluid exert shear forces on each other. The terms F_1 and F_2 represent the x and y components of forces[2] such as gravitational pull and frictional drag at boundaries. For our applications we can take F_1 to be (see [11])

---

[1] In steady motion all partial derivatives with respect to time vanish.

[2] The dimensions of each term in (3.9) and (3.10) are actually force per unit volume.

$$F_1 = -\frac{\hat{f}}{2D}\rho|u|u - \rho g \cos\theta, \tag{3.11}$$

where $\hat{f}$ is another friction factor (cf. equation (3.3)) and D is a quantity having the units of length known as the hydraulic or equivalent diameter.[1] The angle θ is that between the x-axis and the direction of the gravitational vector.

If we integrate (3.9) along a link s_j which is parallel to the x-axis we obtain

$$\Delta p_j = p_P - p_Q = -\int_P^Q F_1\, dx + I, \tag{3.12}$$

where

$$I = \int_P^Q \left[\rho\left(u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}\right) - \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)\right] dx.$$

By the mean value theorem for integrals

$$-\int_P^Q F_1\, dx = \frac{\hat{f}}{2D} L_j \rho|\bar{u}|\bar{u} + \rho g\, L_j \cos\theta, \tag{3.13}$$

where $\bar{u}$ is the velocity at some point on the link.

Therefore, if we define the link mass flow rate to be $w_j = \rho A_j\, \bar{u}$, and let $F_j = fL_j/2D$, then (3.13) becomes

$$-\int_P^Q F_1\, dx = \frac{F_j}{\rho A_j^2}|w_j|w_j + \rho g\, L_j \cos\theta.$$

Consequently it follows from (3.12) that

$$\Delta p_j = \frac{F_j}{\rho A_j^2}|w_j|w_j + \rho g\, L_j \cos\theta + I. \tag{3.14}$$

Comparing this equation with (3.4), we see that they differ only by the integral I. If we accept the Navier-Stokes equations as providing an accurate description of the fluid's motion, then I is the error introduced by using the characteristic (3.5).

---

[1] Except in the case of flow in circular pipes, where D coincides with the pipe diameter, the physical significance of this quantity is ambiguous (see [9], chapter 11, for further discussion).

Let us examine the terms in I. Since

$$u = \frac{dx}{dt}, \quad v = \frac{dy}{dt},$$

we have from the chain rule

$$\frac{d^2x}{dt^2} = \frac{du}{dt} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt} = u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}.$$

But $a = d^2x/dt^2$ is the acceleration of a fluid element in the x-direction. Thus

$$\int_P^Q \rho\left(u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}\right)dx = \int_P^Q \rho a\, dx \tag{3.15}$$

represents the integral of the inertial forces due to the motion of the fluid along the link.

As we have already mentioned, the remaining part of I arises from the shear effects due to the relative motion of fluid elements. If we call (3.15) the inertial effect, then $-\int_P^Q \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)dx$ may be called the viscous effect. Since our characteristic (3.5) assumes that both of these are zero, we expect the error in (3.5) to be significant in situations when either the viscous or inertial effects are comparable to those produced by the friction and elevation.

4. NUMERICAL SOLUTION OF THE NETWORK EQUATIONS

Having reduced the steam generator flow problem to one of solving the network equations, it remains to formulate an algorithm for their solution. We have seen in Section 2 that these equations possess a unique solution. However, because of the nature of the characteristic (3.5), they are necessarily nonlinear. Therefore, a closed form solution--even in the sense of Gauss Elimination--is out of the question and so we employ an iterative method.

The iterative method that we shall consider is called the Nonlinear Gauss-Siedel (NGS) Method, and is a natural extension of the well known method

method for the linear case.[1] We shall formally define it for a general system of ν equations containing ν unknowns.

Suppose that $F_i(x_1, \ldots, x_\nu)$, $i = 1, \ldots, \nu$ are continuous real valued functions of the real variables x_i, $i = 1, \ldots, \nu$. Then

$$F_i(x_1, \ldots, x_\nu) = 0, \quad i = 1, \ldots, \nu \tag{4.1}$$

represents a system of ν equations and any set of real numbers $x_1^*, \ldots, x_\nu^*$ which satisfies (4.1) is called a solution of this system. The NSG method defines a sequence of vector iterates $x^k = (x_1^k, \ldots, x_\nu^k)$, $k = 0, 1, \ldots$ in the following manner:

Step 1. Choose x^o.

Step 2. Suppose that $x^k (k \geqq 0)$ and x_q^{k+1}, $q = 1, \ldots, i-1$ $(1 \leqq i \leqq \nu)$ have been determined.

Step 3. To determine x_i^{k+1}, find s such that

$$F_i(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, s, x_{i+1}^k, \ldots, x_\nu^k) = 0, \tag{4.2}$$

and then let $x_i^{k+1} = s$.

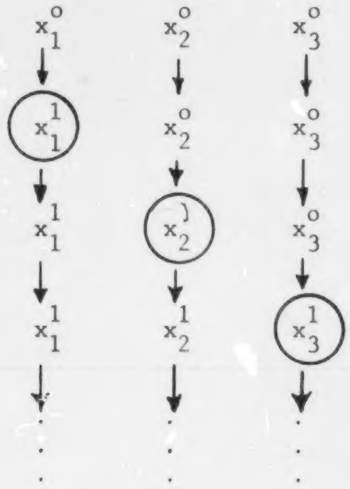
Notice that when $i = 1$ in step 2, the condition on x_q^{k+1}, $q = 1, \ldots, i-1$ is vacuous and x_1^{k+1} is determined in step 3 as the solution s of

$$F_1(s, x_2^k, \ldots, x_\nu^k) = 0.$$

It is clear that the NGS method is well defined, that is the iterates x^k are uniquely determined, if and only if each scalar equation (4.2) has a unique solution s. When this is true, the elements of the NGS sequence are determined in the order illustrated by the tableau for the case $\nu = 3$. Each line of the tableau corresponds to the particular solution of (4.2) which determines the circled variable.

---

[1] The linear method has also been called the "single step method," the "method of successive displacements," and the "Liebmann method." For more details, see [12].



A central question about any iterative method concerns the convergence of its iterates; do there exist numbers x_i^*, $i = 1, \ldots, \nu$ such that

$$\lim_{k\to\infty} x_i^k = x_i^*, \quad i = 1, \ldots, \nu?$$

With regard to the NGS method it is obvious that the iterates not only depend on the functions F_i but also on the initial iterate x^o. Thus, they may converge for some choices of x^o but not for others. In the event that they converge for any choice of x^o, we say that the method is globally convergent. This is an extremely agreeable property for an iterative method to have since it means that we do not have to be particularly concerned about how we start the method to obtain convergence.

Since the functions F_i are continuous, it follows from equation (4.2) that the NGS iterates converge to a solution of (4.1) providing that they converge at all. Therefore, the question of computing a solution to (4.1) by the NGS method hinges on the convergence of its iterates.

That the network equations are a special case of (4.1) may be seen by going back to the form (2.8) and letting

$$(4.3) \qquad F_i(p_1, \ldots, p_\nu) = \sum_{j=1}^{m} a_{ij}\, f_j\Big(\sum_{q=1}^{n} a_{qj}\, p_q\Big), \quad i = 1, \ldots, \nu.$$

At first glance this equation appears to be inconsistent since the left side involves only the first ν pressures whereas the right side involves all $n > \nu$ pressures. However, it must be remembered that nodes $\nu + 1, \ldots, n$ are boundary nodes where the corresponding pressures are presumed to be known. So, in fact, the right side involves only ν unknowns.

The following theorem completely answers the question of how the NGS method behaves relative to network equations. It guarantees--if not the best of all possible worlds--at least a well ordered state of affairs.

Theorem 4.1. If the links of a network $\mathcal{N}$ have admissible characteristics, and if $\mathcal{N}$ has at least one boundary node, then the NGS iterates are well defined, and are globally convergent to the unique solution of the network equations.

The proof of this theorem, although not particularly difficult, is too long to include here. A proof may be found in [13].

Theorem 4.1 asserts that we can use the NGS method to calculate an approximate solution of the network equations, and that this solution can be made to agree with the exact solution to any degree of accuracy provided that we solve (4.2) exactly and iterate long enough. Of course, since (4.2) is a nonlinear scalar equation in the unknown s, it will not be possible in general to solve it exactly. In fact a subsidiary numerical method is usually employed to obtain an approximate solution. This means that in practice we cannot realize the hypotheses of Theorem 4.1. Nevertheless, we expect that if (4.2) is solved accurately, then the resulting approximate NGS iterates will converge to an accurate solution of the network equations.

A second point to be made regarding the practical implementation of the NGS method concerns the question of when to terminate the iterations. A completely satisfactory answer to this question is not known.

Usually the decision to terminate is based on some measure of the difference of successive iterates. For example, if we define the norm of x, $\|x\|$, to be $\left(\sum_{i=1}^{\nu} x_i^2\right)^{1/2}$, then we might stop the iteration at the first value of k such that

$$\|x^k - x^{k-1}\| < \epsilon \, \|x^k\|, \tag{4.4}$$

where $\epsilon > 0$ is some preselected convergence criterion. Note that (4.4) in no way guarantees that the norm of the error, $\|x^k - x^*\|$ is less than ϵ.

A more sophisticated stopping criterion can be formulated by taking into account the rate of convergence of the method. There are several ways to define this rate (see, for example, [14], chapter 9), but in practice it usually must be estimated as the iteration proceeds. The paper [15] contains a stopping criterion which employs the convergence rate.

The idea of a convergence rate serves another useful purpose--it frequently allows us to compare different iterative methods. For instance, we obtain the Nonlinear Successiver Overrelaxation Method from the NGS method simply by letting $x_i^{k+1} = \omega s + (1 - w)x_i^k$ in step 3. Here, $1 \leqq \omega < 2$ is called the relaxation parameter. The NGS method is recovered by letting $\omega = 1$. However, choices of ω significantly greater than 1 frequently give iterates whose convergence is dramatically faster than that of the NGS iterates (see the sample problem in the next section). What is even more remarkable is that for a large class of problems, this is quantitatively predicted by the convergence rates. For more information on the successive overrelaxation method the reader is referred to [12] and [16] for the linear case and [13], [15] for the nonlinear case.

We conclude this section with a final word about the numerical solution of equation (4.2) in the case where the underlying system (4.1) represents the network equations. For (4.3), we see that the i^{th} network equation is

$$\text{(4.5)} \qquad \sum_{j=1}^{m} a_{ij}\, f_j\Big(\sum_{q=1}^{n} a_{qj}\, p_q\Big) = 0,$$

from which it follows that the i^{th} equation (4.2) is

$$\text{(4.6)} \qquad \sum_{j=1}^{m} a_{ij}\, f_j\Big(\sum_{1\leqq q<i} a_{qj}\, p_q^{k+1} + a_{ij}s + \sum_{i<q\leqq\nu} a_{qj}\, p_q^{k} + \sum_{\nu<q\leqq n} a_{qj}\, p_q^{*}\Big) = 0.$$

We recall that a_{ij} is the element in the i^{th} row and j^{th} column of the incidence matrix and the functions $f_j(t)$ define the admissible link characteristics.[1] For the steam generator flow problem they are given explicitly by (3.5).

---

[1] To test his understanding of the nature of (4.6) the reader should convince himself that for the network of Figure 4, the equation which applies when $i = 2$ is $-f_2(p_{14}^* - s) + f_5(s - p_5^k) - f_{16}(p_1^{k+1} - s) + f_{17}(s - p_3^k) = 0$.

Now each term in (4.6) may be written $a_{ij} f_i (a_{ij} s + C^k_{ij})$ where $a_{ij} = 0, \pm 1$ and C^k_{ij} is a constant. Since the characteristics are admissible, it follows that each nonvanishing term of the sum (4.6) is a strictly increasing function of s which changes sign, and so the same is true of the sum. Therefore, there is exactly one value of s which solves (4.6) and we compute this numerically by such methods as bisection, regula falsi, etc.[1]

5. ILLUSTRATIVE PROBLEMS

In this section we pursue the investigation of the flow problem associated with the network of Figure 4. This problem has already been discussed in Section 3 and gives rise to a set of 12 network equations of the form (4.5). These are completely described by the incidence matrix, the data which defines the link characteristics (3.5), and the boundary conditions.

The elements of the incidence matrix may be obtained directly from Figure 4 in an obvious manner. For example, the only nonzero elements of the first row are $a_{11} = -1$, $a_{14} = 1$ and $a_{1,16} = 1$.

To define the checteristics (3.5) we take $\rho = 49 \text{ lb/ft}^3$, $g = 32.2 \text{ ft/sec}^2$ and the remaining data from Table 1. The density we have chosen is approximately the density of water about to boil at a system pressure of 600 psia.

As boundary conditions we assume that $p_{13} = p_{14} = p_{15} = 7600. \text{ pd/ft}^2$ and $p_{16} = p_{17} = p_{18} = 0.0 \text{ pd/ft}^2$.

The problem defined by the above conditions was solved on a computer by the NGS method described in Section 4. The initial pressures were taken to be $p^o_i = 330. \text{ pd/ft}^2$, $i = 1, \ldots, 12$. After 36 iterations the quantity

$$[\sum_{i=1}^{12} (p_i^{36} - p_i^{35})^2]^{1/2}$$ was less than 10^{-4} and the problem was declared to be converged. The pressures and flows thus obtained are given in Tables 2 and 3. The problem was then resolved by the Nonlinear Successive Overrelaxation Method with $\omega = 1.33$. Essentially the same solution was obtained in 19 iterations, illustrating the advantage that can be gained by iterating with $\omega > 1$.

---

[1] See [17], [18] for a description and analysis of these basic methods.

j	L_j(ft)	A_j(ft^2)	F_j	θ_j(deg)
1	.5	.28	$.106 \times 10^{-3}$	0.
2	.5	.21	$.136 \times 10^{-3}$	0.
3	.5	.28	$.106 \times 10^{-3}$	0.
4	1.25	.124	$.53 \times 10^{-3}$	0.
5	1.25	.093	$.7 \times 10^{-3}$	0.
6	1.25	.124	$.53 \times 10^{-3}$	0.
7	.16	.01	$.6 \times 10^{-3}$	0.
8	.16	.01	$.6 \times 10^{-3}$	0.
9	.16	.01	$.6 \times 10^{-3}$	0.
10	1.25	.124	$.53 \times 10^{-3}$	0.
11	1.25	.093	$.7 \times 10^{-3}$	0.
12	1.25	.124	$.53 \times 10^{-3}$	0.
13	.5	.28	$.106 \times 10^{-3}$	0.
14	.5	.21	$.136 \times 10^{-3}$	0.
15	.5	.28	$.106 \times 10^{-3}$	0.
16	.5	.42	$.112 \times 10^{-2}$	90.
17	.5	.42	$.112 \times 10^{-2}$	90.
18	.5	.047	.0105	90.
19	.5	.047	.0105	90.
20	.5	.047	.0105	90.
21	.5	.047	.0105	90.
22	.5	.42	$.112 \times 10^{-2}$	90.
23	.5	.42	$.112 \times 10^{-2}$	90.

Table 1

Node	Pressure[1] (pd/ft^2)	Node	Pressure (pd/ft^2)	Node	Pressure (pd/ft^2)
1	6810.6	2	6810.6	3	6810.6
4	4826.6	5	4820.2	6	4826.6
7	2773.4	8	2779.8	9	2773.4
10	789.41	11	789.41	12	789.41

Table 2

Link	Flow (lb/sec)	Link	Flow (lb/sec)	Link	Flow (lb/sec)
1	136.3	2	90.8	3	136.3
4	129.3	5	104.7	6	129.3
7	121.3	8	120.8	9	121.3
10	129.3	11	104.7	12	129.3
13	136.3	14	90.8	15	136.3
16	6.94	17	-6.95	18	8.08
19	-8.08	20	-8.08	21	8.08
22	-6.95	23	6.93		

Table 3

---

[1] To convert poundals per square foot to the more familiar pounds per square inch multiply by $.216 \times 10^{-3}$. Thus the pressure at node 1 is 1.47 lb/in^2. Note that the total pressure variation over the domain of the problem is less than 1.35 pounds per square inch. This is consistent with our earlier assumption of a constant system pressure (see equation (3.8)).

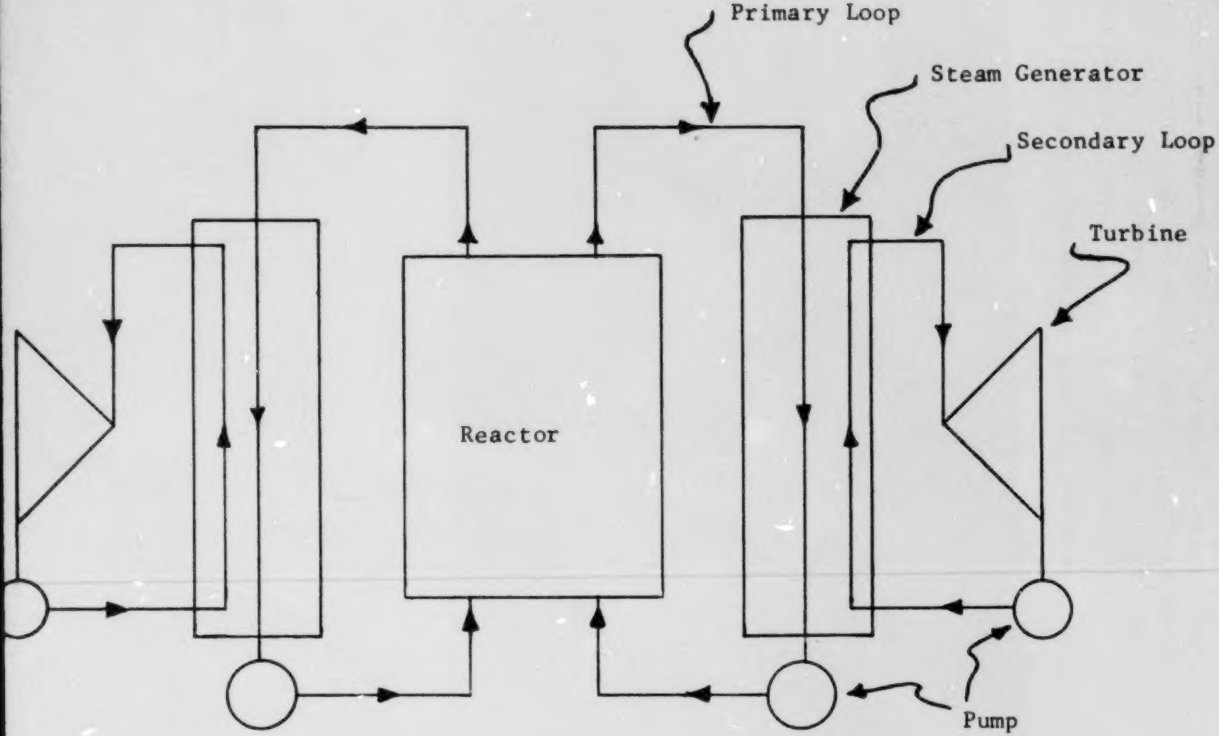

Primary Loop
Steam Generator
Secondary Loop
Turbine
Reactor
Pump

Figure 1

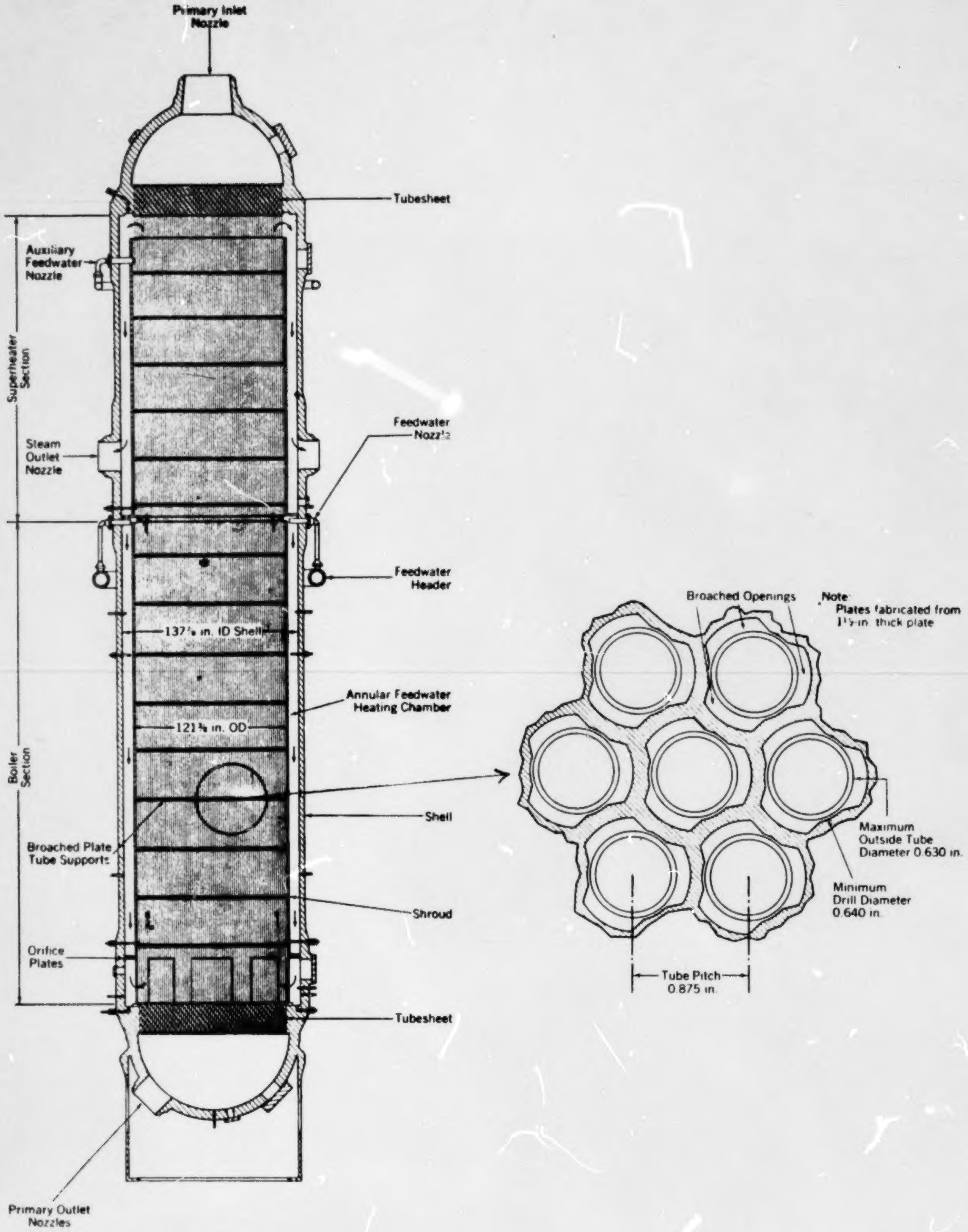

Primary Inlet Nozzle
Tubesheet
Auxiliary Feedwater Nozzle
Superheater Section
Steam Outlet Nozzle
Feedwater Nozzle
Feedwater Header
137⅛ in. ID Shell
Annular Feedwater Heating Chamber
121¾ in. OD
Boiler Section
Shell
Broached Plate Tube Supports
Shroud
Orifice Plates
Tubesheet
Primary Outlet Nozzles
Broached Openings
Note:
Plates fabricated from 1½-in. thick plate
Maximum Outside Tube Diameter 0.630 in.
Minimum Drill Diameter 0.640 in.
Tube Pitch 0.875 in.

Figure 2. Cross section of nuclear once-through steam generator. Reproduced from [1]

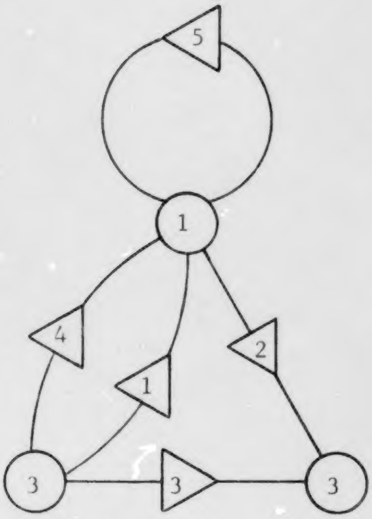


Figure 3

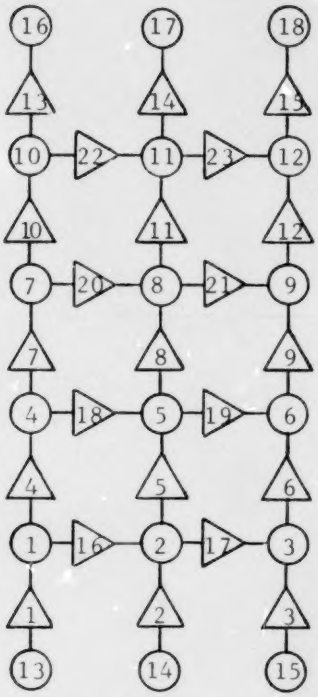

16
17
18
13
14
15
10
22
11
23
12
10
11
12
7
20
8
21
9
7
8
9
4
18
5
19
6
4
5
6
1
16
2
17
3
1
2
3
13
14
15

Figure 4

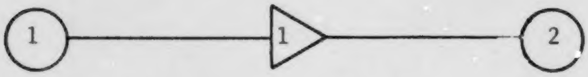


Figure 5

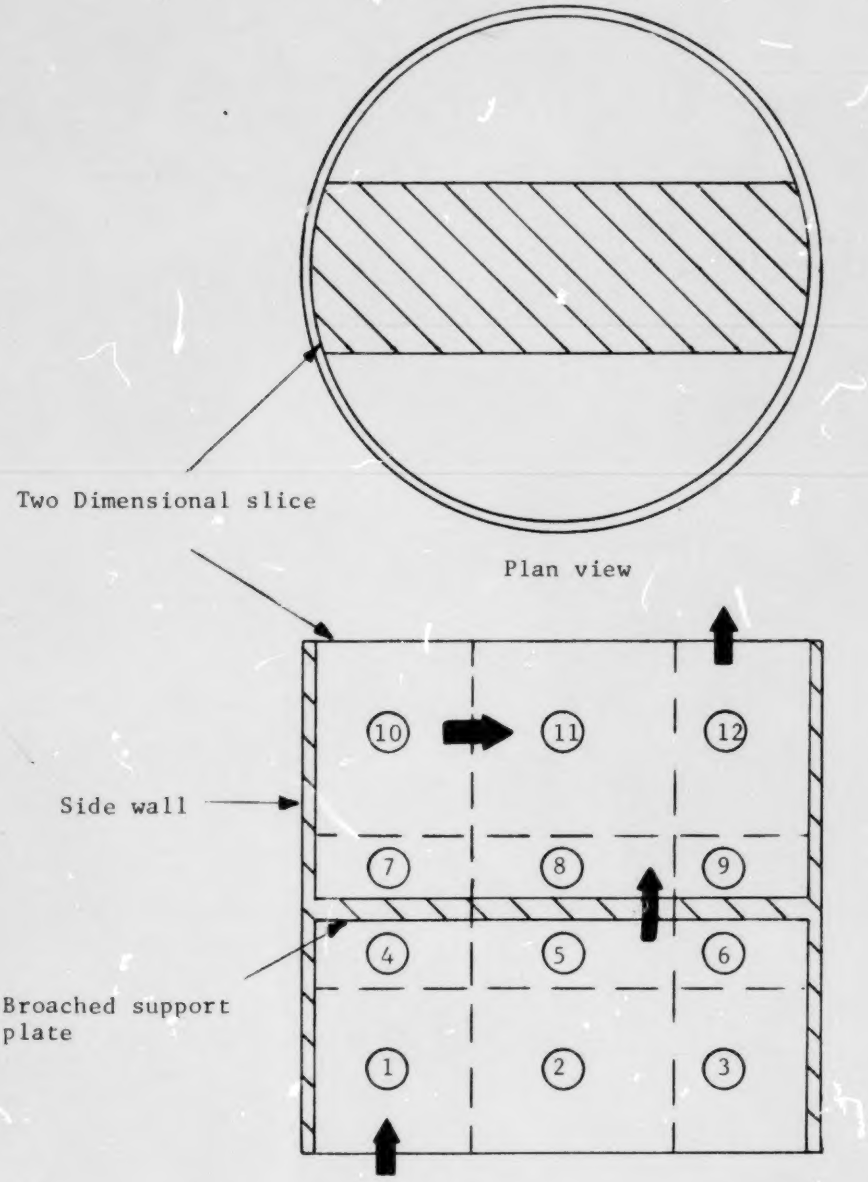

Two Dimensional slice
Plan view
10
11
12
Side wall
7
8
9
4
5
6
Broached support plate
1
2
3

Figure 6

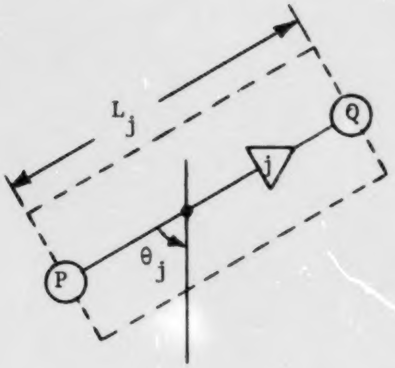

L_j
Q
j
P
θ_j

Figure 7

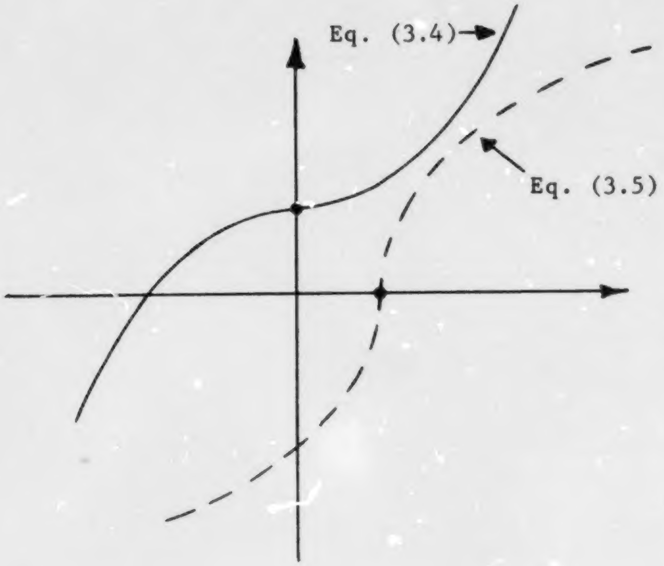


Figure 8

400

Behavior of Deposition Coefficient with Fluid Velocity

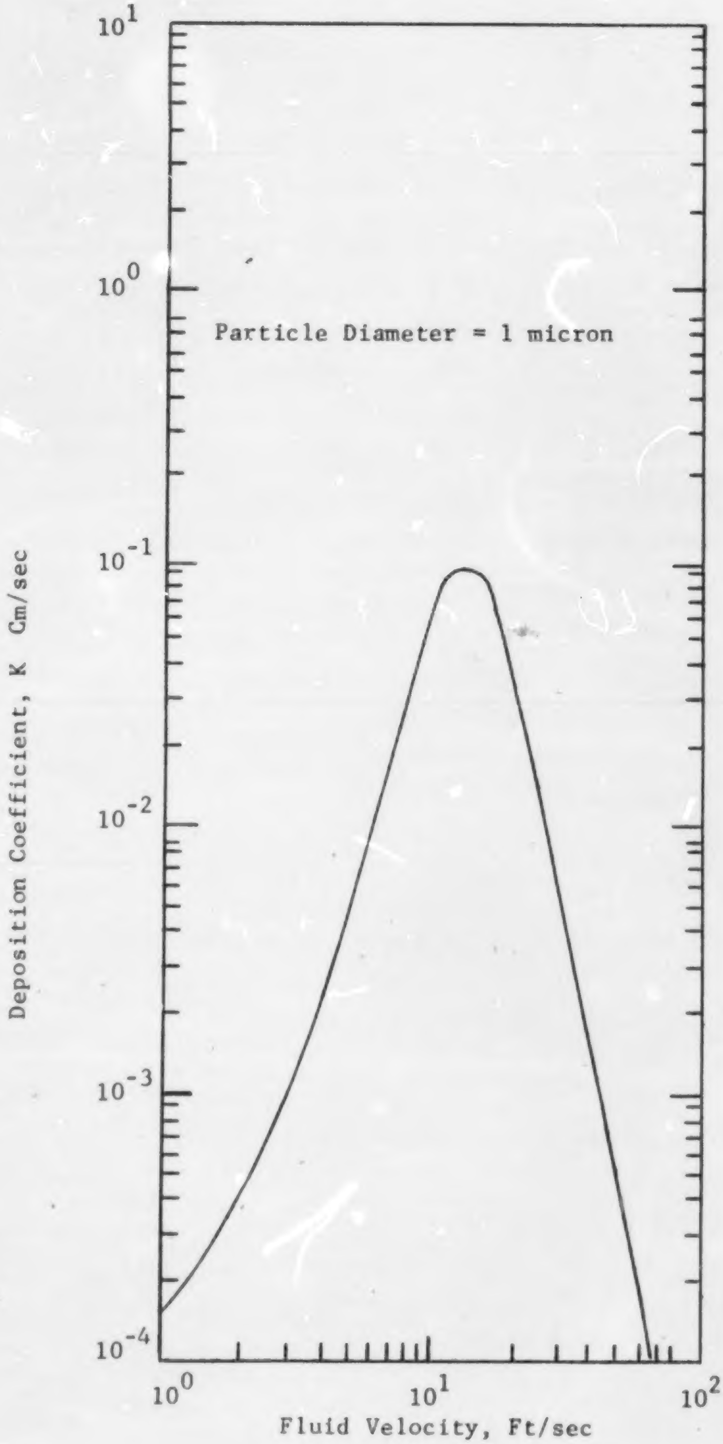


Figure 9. From [20]

REFERENCES

[1] "Steam, Its Generation and Use," Babcock and Wilcox, New York, (1972).

[2] Potts, R. B., Oliver, R. M. "Flows in Transportation Networks," Academic Press, New York (1972).

[3] Ford, L. R., Fulkerson, D. R. "Flows in Networks," Princeton University Press, Princeton, New Jersey (1962).

[4] Berge, C., Ghouila-Houri, A. "Programming, Games and Transportation Networks," Wiley, New York (1965).

[5] Busacker, R. G., Saaty, T. L. "Finite Graphs and Applications," McGraw-Hill, New York (1965).

[6] The Holy Bible, Exodus 14:21.

[7] MacDuffee, C. C. "Vectors and Matrices," Carus Mathematical Monograph Number 7, Mathematical Association of America (1943).

[8] Birkhoff, G. A Variational Principle for Nonlinear Networks, Q. Appli. Math., 21 (1963), pp. 160-162.

[9] Knudsen, J. G. and Katz, D. L. "Fluid Dynamics and Heat Transfer," McGraw-Hill, New York, Toronto, London (1958).

[10] Von Mises, R., Friedrichs, K. O. "Fluid Dynamics," Applied Mathematical Sciences, Vol. 5, Springer-Verlag, New York (1971).

[11] Meyer, J. E. "Hydrodynamic Models for the Treatment of Reactor Thermal Transients," Nucl. Sci. and Eng. 10 (1961), pp. 269-277.

[12] Varga, R. S. "Matrix Iterative Analysis," Prentice-Hall, Englewood Cliffs, New Jersey (1962).

[13] Rheinboldt, W. "On M-Functions and Their Application to Nonlinear Gauss-Seidel Iterations and Network Flows," J. Math. Anal. and Appl., 32 (1970), pp. 274-307.

[14] Ortega, J., Rheinboldt, W. "Iterative Solution of Nonlinear Equations in Several Variables," Academic Press, New York (1970).

[15] Hageman, L. A., Porsching, T. A. "Aspects of Nonlinear Block Successive Overrelaxation," SIAM Journal on Numerical Analysis, (1975).

[16] Young, D. "Iterative Solution of Large Linear Systems," Academic Press, New York (1971).

[17] Ostrowski, A. "Solution of Equations and Systems of Equations," second edition, Academic Press (1966).

[18] Traub, J. "Iterative Methods for the Solution of Equations," Prentice-Hall, Englewood Cliffs, New Jersey (1964).

[19] Beal, S. K. "Deposition of Particles in Turbulent Flow on Channel or Pipe Walls," Nucl. Sci. and Eng. 40 (1970), pp. 1-11.

[20] Beal, S. K. "Prediction of Heat Exchanger Fouling Rates - A Fundamental Approach," Preprint of paper presented at AICHE Meeting, November, 1972.

[21] Blackwell, W. A. "Mathematical Modeling of Physical Networks," Macmillan Company, New York (1968).

Chapter 11
HEAT TRANSFER IN FROZEN SOIL

Gunter H. Meyer
Georgia Institute of Technology

A SHORT GUIDE TO THE MODULE

Section 1 contains a description of the problem to be solved and the motivation for the work reported on below.

Section 2 presents a self-contained formulation of the mathematical model based on elementary heat transfer principles.

Section 3 contains a solution algorithm for the mathematical model. It illustrates the method of lines for partial differential equations, and an initial value solution technique for two point boundary value problems for ordinary differential equations.

Section 4 contains some representative results from an industrial parameter study.

Section 5 presents a mathematical convergence proof. It illustrates use of the maximum principle for differential equations, use of the Ascoli-Arzelà theorem, and the concept of a weak solution for a differential equation.

Prerequisites: For sections 2 and 3 only calculus and some mathematical maturity are needed. No particular knowledge of partial differential equations is required, but previous exposure to ordinary differential equations is helpful. Section 5 requires a strong background in mathematics as is apparent from its contents. This section may not be suitable for undergraduate students.

1. Description of the Problem

Between the Arctic Ocean and the Brooks Range of Alaska lies the North Slope. In this cold and barren tundra oil has been found and considerable oil exploration and production activity is anticipated for the coming years. A good part of this activity will require the erection of engineering structures such as drill rigs, pipelines, work camps and, of course, roads and air fields.

Construction on the North Slope as elsewhere in Arctic and Antarctic regions is complicated by the fact that the gound is in a state of permafrost. Except for an active surface layer of several feet the moisture in the ground stays permanently frozen down to a depth of six hundred feet and more. During most of the year the top layer is also frozen but during the summer thaw the ice in the soil will melt to an average depth of 2 to 3 feet. It is easy to visualize that man-made structures placed on frozen ground will sink into the ground if the ice in the underlying soil is allowed to melt. The problem is rendered even more severe if the structure emits heat, such as a buried pipeline carrying hot oil.

Over many decades of experience with construction in the Arctic both here and abroad, methods have evolved for coping with the problem of permafrost under buildings and roads. One possible avenue is the insulation of the structure from the ground in some form to prevent melting of the ice and the resulting loss of bearing strength. This insulation usually takes the form of a foundation for the structure consisting of sand, gravel, wood, and possibly man-made material arranged in various layers of different thickness. If there is plentiful building material within easy access of the construction site the choice of design may not influence the construction cost unduly. On the other hand, it must be expected that many North Slope construction sites are far from any sources of raw building material. Moreover, it may at times be necessary to fly the material to the construction site since crossing the frozen tundra with tracked vehicles, at this time the main transportation alternative, tears up the ground cover. This vegetation provides a natural insulation and stabilization of the ground; if it is disturbed the soil will melt, the water runoff will wash out gullies which disturb more of the vegetation. A non-reversible cycle will have begun which permanently will scar the countryside.

If indeed building materials have to be transported over long distances construction cost will depend critically on the weights of the various construction components. In order to find the most economical design for the structure there is a need for a fine-tuned method to evaluate whether a given design can meet the requirements placed on it. For structures in permafrost regions the main requirement is prevention of thawing of unconsolidated moisture rich soil and subsiding of the structure into muck. Consequently, a major component of a method of evaluation is the determination of the temperature profile in the ground.

This module will describe the formulation and application of a mathematical model for the simulation of heat flow in a layered medium which has been used extensively by one major oil company for a screening examination of various design alternatives for the construction of roads and airstrips on the North Slope. We shall derive in some detail the equations for the temperature field in a partially frozen medium, present an approximate (eventually numerical) method for its solution, and discuss some representative results and conclusions obtained from this model. These points were an integral part of bringing the model to bear on pressing engineering problems. Finally, to satisfy a mathematician's curiosity we shall discuss the convergence of the approximate solution to the actual solution of the heat flow equation.

Comments:

Some background material on the Arctic and the problems faced by the oil industry may be found in the following three reports. 1) W. S. Ellis, "Will Oil and Tundra Mix?," National Geographic 140 (1971); R. D. Guthrie et al., "North to the Tundra," National Geographic 141 (1972); and B. Keating, "North for Oil," National Geographic 137 (1970).

A reader interested in the technical aspects of permafrost is referred to the Proceedings, Permafrost International Conference, 1963, National Academy of Sciences-National Research Council Publication No. 1287, Washington, D. C.

2. The Model

It is desired to formulate a mathematical model for the heat flow in and under such structures as roads, airstrips and building foundations resting on a layered soil in a state of permafrost. One may think, for example, of the construction of an airfield. The natural ground may consist of several layers of frozen sand and silt onto which several man-made layers of sand, gravel, possibly an insulator like styrofoam, and concrete or asphalt have been placed to distribute the expected loads and to prevent the frozen ground from melting. It can be expected that during the summer thaw moisture will enter into the man-made top layers which subsequently will participate in the seasonal freezing process and, as will be shown, materially influence the insulating properties of the top layers.

It is possible to construct a full three dimensional model for heat flow under the structure. However, the resulting equations will be quite complicated and difficult to solve. On the other hand, many years of experience have taught engineers that the thermal effects at the edge of the structure will be felt only within three or four feet of the edges. Since the structure may well extend over 30 or more feet in each direction the temperature field under the major portion of the structure will depend only on depth. For this reason the decision is made to ignore edge effects and to model the temperature field as a function of time and depth only. In this setting the structure foundation becomes an "infinite" slab.

It must be emphasized that this approximation is based on past experience; should the results obtained from such a simplified model not be consistent with experience the model must be improved. It is important in any study of this kind to remain aware of the limitations built into the model.

One distinguishes between three different modes of heat transmission, called radiation, convection and conduction. On the scale of the individual sand grains and pebbles all three modes are undoubtedly present in a water logged soil. However, experiments have shown that the overall transmission of heat in such a medium can adequately be described by a pure heat conduction model. We shall now present this model.

Let us begin by considering one layer of moisture free homogeneous soil, say sand. The top of the layer is thought to be exposed to the atmosphere and the the bottom for the time being, is held at a fixed temperature. In order to determine the temperature profile between top and bottom a mathematical model for the prediction of heat flow in the layer can be formulated. As stated above, it will be assumed that heat flow in this layer is due to conduction which means that the following three empirical laws (or experimental facts) govern the flow of heat in the layer [1]

a) heat flows in the direction of decreasing temperature

b) the quantity of heat gained or lost by a body during a temperature change is proportional to the mass of the body and to the temperature change

c) the rate of heat flow through an area is proportional to the area and to the temperature gradient normal to the area.

If u is the average temperature of the body, Q its heat content, and ρ its density then with reference to Figure 1 the second of these laws can be expressed as

$$\Delta Q = c\Delta u \, \rho\Delta x A \tag{2.1}$$

The proportionality constant c is known as the specific heat of the material; $\rho\Delta x A$ is, of course, the mass of the control volume. The third law is known as Fourier's law of heat conduction. If we assume that the heat flow in a neighborhood of the point of interest in the layer is strictly vertical then we can write for the heat loss ΔQ per time span Δt at the mean time t through an area A at depth x in the direction of increasing x

$$\frac{\Delta Q}{\Delta t} = -kA\frac{\partial u}{\partial x}(x,t) \tag{2.2}$$

where the constant k of proportionaly is known as the conductivity of the material. Consider now an energy balance for the control volume in Figure 1. The amount of heat flowing through the face at $x+\Delta x$ minus that flowing through the area at x must be equal to the amount of heat lost per time span Δt, or

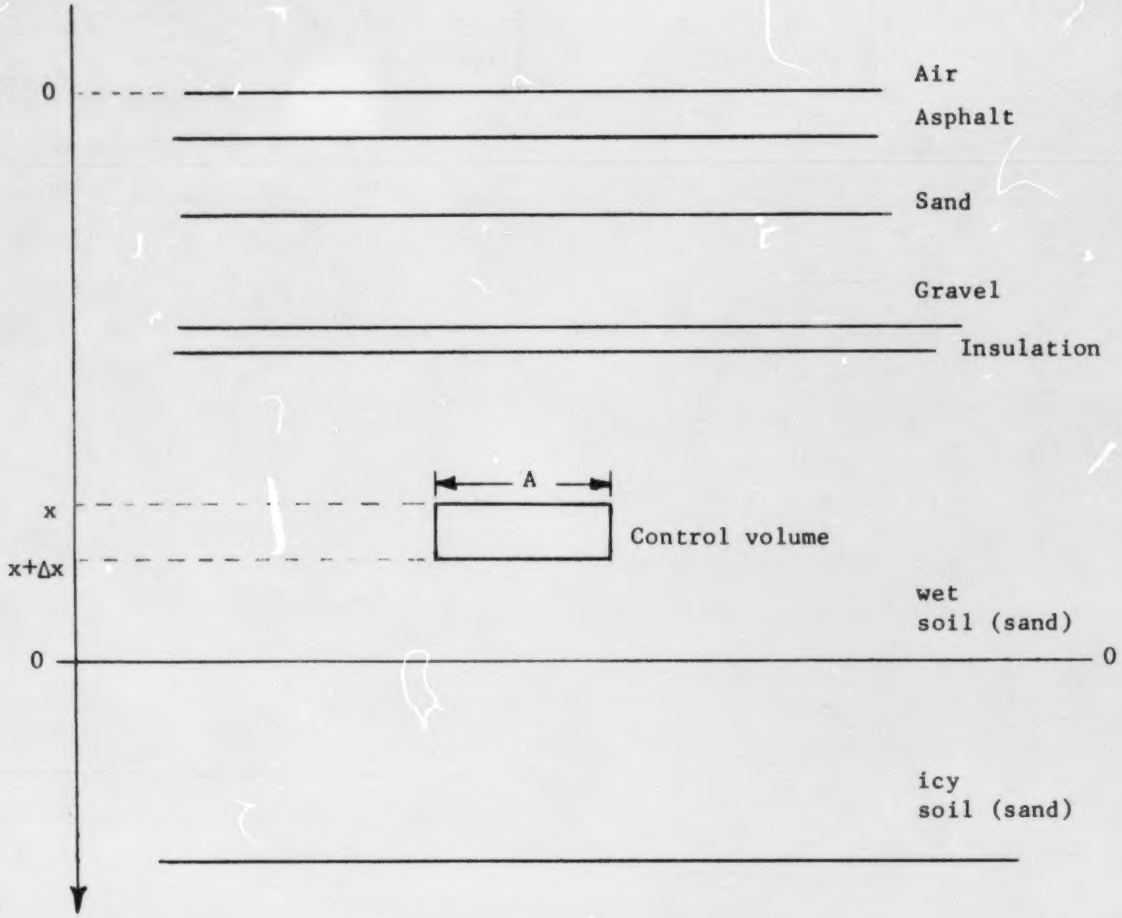

0
Air
Asphalt
Sand
Gravel
Insulation
A
x
x+Δx
Control volume
wet
soil (sand)
0
0
icy
soil (sand)

Figure 1. Typical arrangement of 4 structural layers on a partially frozen soil

$$-kA\frac{\partial u}{\partial x}(x+\Delta x,t) + kA\frac{\partial u}{\partial x}(x,t) = -c\rho\ A\ \Delta x\ \frac{\Delta u}{\Delta t}$$

Assuming, as is reasonable for the system at hand, that the temperature varies smoothly with time and space we can take the limit as $\Delta x, \Delta t \to 0$ and obtain Fourier's heat conduction equation for the temperature u at any point (x,t) in space and time

$$k\frac{\partial^2 u}{\partial x^2}(x,t) = c\rho\frac{\partial u}{\partial t}(x,t). \tag{2.3}$$

At $x = X$ the temperature is held fixed at some value u_X, i.e.,

$$u(X,t) = u_X. \tag{2.4}$$

The proper boundary condition at $x = 0$ is not quite as easy to formulate since the soil and surrounding air may well be at different temperatures. We shall say more about this later on. At this point we shall simply assume that the temperature of the top soil is a specified function of time, say

$$u(0,t) = \alpha(t). \tag{2.5}$$

Finally, we need an initial temperature profile for the soil at some starting time t_0 of our observation. For convenience we shall choose $t_0 = 0$ and write

$$u(x,0) = u_0(x). \tag{2.6}$$

The actual form of u_0 may vary from problem to problem.

The equations (2.3 - 2.6) define a well posed initial/boundary value problem. Its solution, for reasonable (say continuous) data is known to exist, and experience has shown that it quite well predicts the temperature field within a homogeneous medium subject to one dimensional heat flow. Of course, in order to use this model for engineering calculations one needs to know the functions $\alpha(t), u_0(x)$ and the physical constants k, c, and ρ for the layer. Such data either already are, or are rapidly becoming available from geological, climatological and soil studies.

So far we have assumed that the layer was homogeneous. Let us now go one step further and assume that the layer still is homogeneous, but that it contains moisture in the form of water and ice. For definiteness, let us choose the layer of sand of Figure 1 but suppose that at a given time t_0 the sand is wet above the line 00, while below it the sand is frozen (this requires, of course, that $\alpha(t_0) > 32^\circ F$, $u_X < 32^\circ F$). It is now convenient to consider the sand as two layers. For layer I the heat equation (2.3) and the boundary condition (2.5) is valid, where k, c, and ρ are values determined for a wet sand. Below the freezing line the heat equation (2.3) is valid when values for a frozen sand are chosen for k, c, and ρ. Of course, the boundary condition (2.4) remains in effect. In order to compute the temperature profile at future times we need to link the temperatures and heat fluxes in the wet and frozen sand at the interface AB, the so-called free interface. The temperature condition is simple. Both the bottom of the wet sand and the top of the frozen sand must be exactly at the freezing temperature of water (taken to be $32^\circ F$) so that

$$u_{SW}(s(t),t) = u_{SI}(s(t),t) = 32 \tag{2.7}$$

where SW and SI denote the wet and icy layer of sand and where $s(t)$ denotes the location of the free interface at time t. The free interface, of course, will move in time. Suppose that in a time span Δt the interface has moved downward by a distance Δs. Then the amount of heat used up in melting a block $\Delta s \cdot A$ of frozen sand is given by

$$Q = \rho_{SI}\lambda\Delta s \cdot A$$

where ρ_{SI} is the density of the frozen sand and λ is the latent heat of fusion of the frozen sand. The heat required for melting is provided by the heat flux out of the wet sand minus the heat lost by transmission into the icy sand. Thus

$$Ak_{SI}\frac{\partial u}{\partial x}SI - Ak_{SW}\frac{\partial u}{\partial x}SW = \rho_{SI}\lambda\frac{\Delta s}{\Delta t}A.$$

Letting $\Delta t \to 0$ we find the second free interface condition as

$$k_{SI}\frac{\partial u}{\partial x}SI - k_{SW}\frac{\partial u}{\partial x}SW = \rho_{SI}\lambda s'(t). \tag{2.8}$$

Notice from this development that ρ_{SI} must be replaced by ρ_{SW} if refreezing occurs at the free interface. Thus in order to find the temperature distribution in a partially frozen homogeneous soil we need to solve simultaneously for the solution $\{u_{SW}, u_{SI}, s(t)\}$ of the two heat equations valid above and below $s(t)$ subject to the given boundary and interface conditions.

The discussion of our model is not yet complete. The structures we have in mind consist of several homogeneous but dissimilar layers which in turn may rest on a layered soil. For each homogeneous layer the above development holds. At the boundary between two layers we shall assume perfect thermal contact so that the temperature and the heat influx is continuous at this point. For example, if at depth $x = x_k$ there is an interface between a sand and underlying gravel layer then

$$\lim_{\varepsilon\to 0}\left[u(x_k+\varepsilon,t) - u(x_k-\varepsilon,t)\right] = 0$$

$$\lim_{\varepsilon\to 0}\left[k_g\frac{\partial u}{\partial x}(x_k+\varepsilon,t) - k_s\frac{\partial u}{\partial x}(x_k-\varepsilon,t)\right] = 0$$

where the indices s and g denote sand and gravel. Note that the actual values for the conductives k_g and k_s depend on whether the sand and gravel at location x_k are wet or frozen.

Let us summarize the equations for the mathematical model of heat conduction in a layered medium. We shall suppose that there are altogether M fixed dissimilar layers, where the ith layer extends from depth x_{i-1} to x_i. The thermal parameters are as follows.

$k_i^f (k_i^w)$ = thermal conductivity of the ith layer when frozen (wet).

ρ_i = density of the ith layer (assumed equal for wet and frozen soil--not a serious restriction).

$c_i^f (c_i^w)$ = heat capacity of the ith layer when frozen (wet).

λ_i = latent heat of fussion for the ith layer.

The object is to find the function

$u(x,t)$ = temperature at x at time t.

and the function

$s(t)$ = interface between the frozen and wet ground.

As additional data we specify

$\alpha(t)$ = surface temperature of the top layer, i.e. at $x = 0$.

u_X = constant temperature at the bottom of the last layer, i.e., at $x = x_M$, always held below freezing.

$u_0(x)$ = temperature profile in the layers at the start of the computation.

With this notation the temperature profile in the layered medium can be determined from the following free boundary problem:

(2.9a) $$k_i^\gamma \frac{\partial^2 u}{\partial x^2} - \rho_i c_i^\gamma \frac{\partial u}{\partial x} = 0,$$

with the boundary conditions at the surface

(2.9b) $$u(0,t) = \alpha(t),$$

the heat flow conditions at the fixed interfaces

(2.9c) $$\lim_{\varepsilon \to 0} [u(x_i - \varepsilon, t) - u(x_i + \varepsilon, t)] = 0$$

$$\lim_{\varepsilon \to 0} \left[k_i^{\gamma} \frac{\partial u}{\partial x} (x_i - \varepsilon, t) - k_{i+1}^{\gamma} u(x_i + \varepsilon, t) \right] = 0 \tag{2.9d}$$

and the bottom boundary condition

$$u(X,t) = u_X , \tag{2.9e}$$

where $\gamma = w$ if the ground is wet, $x \varepsilon (0,s(t))$, and where $\gamma = f$ if the ground is frozen, $x \varepsilon (s(t),X)$. The position of $s(t)$ is determined from the free interface condition

$$\lim_{\varepsilon \to 0} u(s(t) + \varepsilon, t) = \lim_{\varepsilon \to 0} u(s(t) - \varepsilon, t) = 32 \tag{2.9f}$$

$$\lim_{\varepsilon \to 0} \left[k_j^{f} \frac{\partial u}{\partial x} (s(t) + \varepsilon, t) - k_j^{w} \frac{\partial u}{\partial x} (s(t) - \varepsilon, t) \right] = \rho_j \lambda_j \frac{ds}{dt} \tag{2.9g}$$

where j is the index of the layer which contains $s(t)$. Finally, an initial temperature distribution at the beginning of the computation, say at $t = 0$, is assumed

$$u(x,0) = u_0(x). \tag{2.9h}$$

The point s_0 where $u_0(x) = 32$ is the initial location of free interface so that

$$s(0) = s_0. \tag{2.9i}$$

We shall always assume that s_0 is uniquely determined by $u_0(x)$. It is true, however, that the importance of $u_0(x)$ and s_0 on the solution $\{u(x,t),s(t)\}$ diminishes rapidly as the system evolves with time.

At first glance the equations (2.9) look forbidding; however, with a little effort some regularity will become apparent. Pick a point with coordinate (depth) x under the structure. This point will fall into one particular layer of material, say the ith layer (which may be gravel, soil, or whatever other materials are present). The temperature at time t at this point is $u(x,t)$. It is either above freezing or below. If $u(x,t) > 32^{\circ}$ then the moisture at this point is present as water and thermal parameters are indexed by $\gamma = w$. If $u(x,t) < 32^{\circ}$ the moisture is present as ice and the thermal parameters are indexed by $\gamma = f$. The dividing line between the wet and frozen material lies at depth $s(t)$ which falls into some layer, say the jth layer. The movement of this line is given by (2.9g) and must be determined together with the temperatures $u(x,t)$ for all M layers.

The problem (2.9a-i) is today commonly known as a two phase Stefan problem in honor of the Austrian physicist J. Stefan (1835-1893) who used such models to study ice-water systems. Problems of this type, however, occur quite frequently in a variety of applications and not just heat transfer with change of phase. They arise in chemical reaction, biological diffusion, visco-plastic diffusion and even stellar evolution [2], [3]. The development that follows in the succeeding sections, particularly the solution algorithms can be readily modified to accomodate many problems with similar mathematical structure.

Suggested Exercises:

1) Derive Fourier's heat conduction equation from Fourier's law of conduction for variable conductivity, i.e., $k = k(x,t)$.
Can you think of a thermal model where a variable k may occur?

2) Derive the heat equation in three space dimensions. Use either the divergence theorem for arbitrary control volumes or consider a cube as control volume.

3) Suppose water is in contact with a warm wall at $x = 0$ and at $s(t)$ with a slab of ice held exactly at 32°F at all times.
Find a free boundary problem describing the temperature in the water.
Is this model realistic for an ice and water system?

References:

1. L. R. Ingersoll, O. J. Zobel and A. C. Ingersoll, Heat Conduction with Engineering and Geological Applications, McGraw-Hill, New York, 1948.

2. J. Ockendon and W. Hodgkins, eds., Moving Boundary Problems in Heat Flow and Diffusion, Clarendon Press, Oxford, 1975.

3. L. Rubinstein, The Stefan Problem, Transl. Math. Monog. 27, A. Solomon, transl., Amer. Math. Soc., Providence, R. I., 1971.

3. A Solution Algorithm

In an industrial environment the folumlation of the mathematical model and the choice of the method of solution, either exact or numerical, are of paramount importance. In general, little leeway, other than some standard linearizing assumptions, exists in formulating the model because the laws of physics have to be obeyed. On the other hand, if the resulting equations do not have a readily obtainable exact solution, considerable latitude exists in the choice of mathematics for solving them approximately. The full interface problem (2.9) with a time varying surface temperature $\alpha(t)$ does not exactly lend itself to a closed form solution in terms of Green's and Neumann's functions which are often used for elementary calculations, so that one usually resorts to a numerical solution technique. The following approach has proved to be very efficient.

First of all the so-called method of lines is employed to replace the heat equation by a sequence of ordinary differential equations. Thus, suppose that we need the solution of our freezing-melting problem over the time interval $[0,T]$. Let $N > 0$ be an integer and define the partition $\{0 = t_o < t_1 < \ldots < t_N = T\}$ of the interval $[0,T]$ where $\Delta t = \frac{T}{N}$, $t_n = n\Delta t$. Replace the partial derivative u_t at $t = t_n$ and the front velocity $\dot{s}(t)$ by the backward difference quotients

$$u_t(x,t_n) \approx \frac{u(x,t_n) - u(x,t_{n-1})}{\Delta t} \equiv \frac{u_n(x) - u_{n-1}(x)}{\Delta t}$$

$$\dot{s}(t_n) \approx \frac{s(t_n) - s(t_{n-1})}{\Delta t} \equiv \frac{s_n - s_{n-1}}{\Delta t}$$

where $u_n(x)$ and s_n are meant to denote the solution as time $t = t_n$. Then at each time level t_n the equations (2.9) can be approximated by the following free interface problem for ordinary differential equations

(3.1) $$k_i^{\gamma}(u_n)'' - \frac{\rho_i c_i^{\gamma}}{\Delta t}[u_n - u_{n-1}(x)] = 0, \quad (u_n)'' \equiv \frac{d^2 u_n}{dx^2}$$

(3.2) $$u_n(0) = \alpha(n\Delta t) \equiv \alpha_n$$

(3.3) $$\lim_{\epsilon \to 0} [u_n(x_i + \epsilon) - u_n(x_i - \epsilon)] = 0$$

$n = 1, \ldots, N$

$i = 1, \ldots, M$

(3.4) $$\lim_{\epsilon\to 0} [k_i^{\gamma} u_n'(x_i - \epsilon) - k_{i+1}^{\gamma} u_n (x_i + \epsilon)] = 0$$

(3.5) $$u_n(X) = u_X$$

(3.6) $$\lim_{\epsilon\to 0} u_n(s_n + \epsilon) = \lim_{\epsilon\to 0} u_n(s_n - \epsilon) = 32$$

(3.7) $$\lambda_j \rho_j \frac{s_n - s_{n-1}}{\quad} - \lim_{\epsilon\to 0} [k_j^f u_n'(s_n + \epsilon) - k_j^w u_n'(s_n - \epsilon)] = 0.$$

where again $\gamma = w$ if $u_n \geqq 32$, $\gamma = f$ if $u_n < 32$. Since $u_o(x)$ and s_o are given these problems must be solved successively for $n = 1, 2,$ etc., for the temperature u_n and the interface s_n.

In the terminology of ordinary differential equations the system (3.1-7) describes an interface problem with M-1 (interior) fixed interfaces and one free interface, the location of the freezing melting isotherm. In general fixed interface problems for linear equations are fairly easy to solve, but free interface problems are not because the problem is no longer linear due to the presence of the unknown interface. The solution algorithm proposed here is based on a conversion to initial value problems. To make the ideas precise consider for the moment the simple two point boundary value problem

(3.8) $$\begin{aligned} u' &= Au + Bv + F(x) \qquad & u(0) &= a \\ v' &= Cu + Dv + G(x) \qquad & u(X) &= b \end{aligned}$$

where F and G are assumed to be continuous.

A possible way of solving (3.8) is to assume an initial value $v(0) = r$ and to integrate the differential equations for u and v over $[0,X]$. Since (3.8) is a linear inhomogeneous system this integration can always be carried out. In fact, the solution may be written as

$$\begin{aligned} u(x) &= \varphi_1(x)r + \psi_1(x) \\ v(x) &= \varphi_2(x)r + \psi_2(x) \end{aligned}$$

where φ_i and ψ_i are determined from the variation of constants procedure. If the second equation is used to eliminate r from the first, then it is seen that regardless of the initial condition r the functions u and v are related through the transformation

(3.9) $$u(x) = U(x)v(x) + w(x)$$

which is known as the Riccati transformation. The functions U and w can be computed directly without going through the variation of constants procedure. Since $u(0) = a$ regardless of what we choose for v it follows that $U(0) = 0$ and $w(0) = a$. To satisfy $u(X) = b$ we must choose

$$b = U(X)v(X) + w(X)$$

or

$$v(X) = \frac{b-w(X)}{U(X)} \tag{3.10}$$

The defining equations for U and w can be obtained from the variation of constants approach described in great detail in the reference listed below. Here we shall merely verify the validity of the appropriate equations.

Let U and w be solutions of

$$U' = B + AU - DU - CU^2, \; U(0) = 0 \tag{3.11a}$$

$$w' = [A - CU(x)]w - U(x)\, G(x) + F(a), \; w(0) = a. \tag{3.11b}$$

These are two initial value problems which can be solved at least in a neighborhood of the initial point $x = 0$. Suppose that $\{U,w\}$ exist over $[0,X]$. Let $\hat{v}$ be given by (3.10), i.e.,

$$\hat{v} = \frac{b-w(X)}{U(X)}$$

and let v be solution over $[0,X]$ of the linear equation

$$v' = [CU(x) + D]v + Cw(x) + G(x), \; v(X) = \hat{v}. \tag{3.12}$$

Then the pair

$$\{u(x) = U(x)v(x) + w(x), \; v(x)\} \tag{3.13}$$

is a solution of (3.8). Indeed, we see that

$$u(0) = U(0)v(0) + w(0) = w(0) = a, \quad u(X) = U(X)\hat{v} + w(X) = b$$

and that $v' = Cu + Dv + G(x)$; moreover

$$u' = U'v + Uv' + w' = [B + AU - DU - CU^2]v + U[CU+D]v + UCw + UG + Aw - UCw - UG + F = A(Uv+w) + Bv + F$$

$= Au + Bv + F(x)$. The equations (3.11a,b), (3.12) and (3.13) are commonly known as the invariant imbedding equations for the boundary value problem (3.8).

Suppose next that a fixed interface is present at $x = L$. Let us write the problem as

$$\left.\begin{aligned} u_i &= A_i u_i + B_i v_i + F_i(x) \\ v_i &= C_i u_i + D_i v_i + G_i(x) \end{aligned}\right. \qquad i = 1, 2 \tag{3.14}$$

(3.15) $u_1(0) = a,\ u_2(X) = b,\ u_1(L) = u_2(L),\ k_1v_1(L) = k_2v_2(L)$

where the subscript 1 refers to the solution $\{u_1, v_1\}$ over $[0,L]$, and where 2 designates the solution over $[L,X]$.

We have seen above that over each subinterval the solutions can be represented by (3.13), where U_i and w_i are given by the differential equations (3.11a) and (3.11b) and where $\hat{v}_2$ is given by (3.10), all functions being properly subscripted with 1 for the interval $[0,L]$ and 2 for $[L,X]$. Missing still are the initial values for $U_2(L)$, $w_2(L)$ and $v_1(L)$. From the fixed interface condition we obtain

$$U_1(L)v_1(L) + w_1(L) = U_2(L)v_2(L) + w_2(L)$$

$$k_1v_1(L) = k_2v_2(L)$$

We do not yet know $v_1(L)$; hence these two equations must be solved for $U_2(L)$ and $w_2(L)$ such that the above relations hold for arbitrary $v_1(L)$.

Notice that if we set

$$U_2(L) = U_1(L)\frac{k_2}{k_1},\ w_2(L) = w_1(L),\ v_1(L) = \frac{k_2}{k_1}v_2(L)$$

(3.16) $w_2(L) = w_1(L)$

then the functions

$\{u_1(x) \equiv U_1(x)v_1(x) + w_1(x),\ v_1(x)\}$ and $\{u_2(x) \equiv U_2(x)v_2(x) + w_2(x)\}$

solve the fixed interface problem.

Finally, let us assume that the boundary $x = X$ is not fixed but free, (i.e., undetermined) but that also a condition on $v(X)$ is specified. To be specific let us suppose that we wish to solve (3.14) subject to

(3.17) $u_1(0) = a, u_1(L) = u_2(L), k_1v_1(L) = k_2v_2(L),$

$$u_2(s) = 0,\ v_2(s) = g(s)$$

where s denotes the (unknown) location of the free boundary and g is a given function over $[L,\infty]$. As outlined above, over $[0,L]$ and $[L,s]$ we have available the representation

$$u_1(x) = U_1(x)v_1(x) + w_1(x)$$

$$u_2(x) = U_2(x)v_2(x) + w_2(x)$$

where U_i and w_i, $i = 1,2$, are known functions, while v_2 and consequently v_1 can be found only after s is known. However, the boundary conditions at s require that s be chosen such that

$$0 = U_2(s)g(s) + w_2(s).$$

Thus consider the functional $\varphi(x) \equiv U_2(x)g(x) + w_2(x)$. Since U_2 and w_2 are assumed to be known (or computed) over $[L,\infty]$ we simply need to find a root s of $\varphi(x) = 0$. If such an s can be found then we integrate (3.12) for $i = 2$ subject to $v(s) = g(s)$ backward over $[L,s]$ and continue with v_1 subject to $v_1(L) = \frac{k_2 v_2(L)}{k_1}$ over $[0,L]$ in order to obtain the complete solution $\{u_i, v_i\}$, $i = 1, 2$. For easy reference we shall refer to the integration of U_i and w_i as the "forward sweep" and to that of v_i as the backward sweep. Thus the free boundary problem is solvable by completing the forward sweep, by determining the free boundary and by carrying out the backward sweep. This approach will now be adapted to the Stefan problem for a layered soil.

Starting with the initial data $u(x,0) = u_0(x)$ and $s(0) = s_0$ we shall advance the solution from time to time. Thus, let us suggest that the temperature profile $u_{n-1}(x)$ and the freezing melting isotherm s_{n-1} are known at time $t_{n-1} = (n-1)\Delta t$ and that we are to find u_n and s_n. For ease of notation let us suppress the subscript n denoting time and instead introduce the subscript i denoting the temperature in the ith layer (at time t_n). In addition it will be helpful on occasion to specifically indicate whether the temperature is above or below freezing. We shall do so with the superscript w for wet and f for frozen ground. For example, with this notation the interface condition (3.6) (at $t = t_n$) can be written as

$$u^f(s) = u^w(s) = 32$$

(The reader is cautioned to distinguish between the function $w_i(x)$ arising in the Riccati transformation, and the superscript w denoting the wet phase.)

In order to cast the problem (3.1-7) into a form similar to (3.8) we shall choose instead of (3.1) the equivalent first order system

(3.18)
$$u_i' = v_i$$
$$v_i = \frac{c_i^\gamma \rho_i}{k_i^\gamma \Delta t}[u_i - u_{n-1}(x)] \equiv \eta_i^\gamma [u_i - u_{n-1}(x)]$$

where $\eta_i^\gamma = \dfrac{c_i^\gamma \rho_i}{k_i^\gamma \Delta t}$ and where $i = 1, \ldots, M$.

The fixed boundary and interface conditions are

$$\begin{aligned} u_1(0) &= \alpha(n\Delta t) \\ u_i(x_i) &= u_{i+1}(x_i) \\ k_i^\gamma v_i(x_i) &= k_{i+1}^\gamma \, v_i(x_{i+1}) \\ u_M(X) &= u_X \end{aligned} \qquad i = 1, \ldots, M\text{-}1 \tag{3.20}$$

In addition we have the free interface conditions

$$u_j^f(s) = u_j^w(s) = 32 \tag{3.23}$$

$$k_j^f v_j^f(s) - k_j^w v_j^w(s) = \lambda_j \rho_j \, \frac{s - s_{n-1}}{\Delta t} \tag{3.24}$$

where j is the index of the layer to which s belongs at $t = t_n$.

The solution algorithm outlined above is readily adapted to this problem. Suppose for the moment that $\alpha(n\Delta t) > 32$ so that a water phase is present. Then we carry out the forward sweep. We know that u_i^w and v_i^w are related through the Riccati transformation

$$u_i^w = U_i(x) v_i^w + w_i(x) \tag{3.25}$$

where U_i and w_i are found successively from the initial value problems

$$U_1' = 1 - \eta_i^w U_i^2, \quad U_1(0) = 0, \quad U_{i+1}(x_i) = \frac{k_{i+1}^w U_i(x_i)}{k_i^w} \tag{3.26}$$

$$w_i' = -\eta_i^w U_i(x)[w_i - u_{n-1}(x)], \quad w_1(0) = \alpha(n\Delta t), \quad w_i(x_i) = w_{i+1}(x_i). \tag{3.27}$$

Similarly, the solutions u_i^f and v_i^f for the ice phase are related through a Riccati transformation which we shall write as

$$u_i^f = R_i(x) v_i^f + z_i, \tag{3.28}$$

where R_i and z_i are determined from the "forward sweep"

$$R_i' = 1 - \eta_i^f R_i^2, \quad R_M(X) = 0, \quad R_{i-1}(x_{i-1}) = \frac{k_{i-1}^f R_i(x_{i-1})}{k_i^f} \tag{3.29}$$

$$z_i' = -\eta_i^f R_i(x)[z_i - u_{n-1}(x)] \quad z_M(X) = u_X, \quad z_{i-1}(x_{i-1}) = z_i(x_{i-1}). \tag{3.30}$$

(Note that this forward sweep carries us through the layers from the lower horizon X to the surface so that these equations are integrated in the direction of decreasing x.) It now remains to find the free interface. If at an arbitrary point x in the jth layer the temperature of both phases is equal to 32 then the corresponding gradients v_i are found from (3.25) and (3.28) as

$$v_j^w(x) = \frac{32 - w_j(x)}{U_j(x)}, \quad v_j^f = \frac{32 - z_j(x)}{R_j(x)} .$$

If this representation is substituted into the interface condition (3.23) we find that s must be a root of the functional (at the nth time level)

$$(3.31) \qquad \varphi(x) \equiv \frac{32 - k_j^f z_j(x)}{R_j(x)} + \frac{k_j^w j_j(x) - 32}{U_j(x)} - \lambda_j \rho_j \frac{x - s_{n-1}}{\Delta t} = 0.$$

Once such a root has been found we can complete the backward sweep by integrating (backward over $[0,s]$)

$$(3.32) \qquad v_i^{w'} = \eta_i^w U_i(x) v_i + \eta_i^w [w_i(x) - u_{n-1}(x)], \quad v_i^w(s) = \frac{32 - w_i(s)}{u_i(s)}$$

and (forward over $[s,x_m]$)

$$(3.33) \qquad v_i^{f'} = \eta_i^f R_i(x) v_i^f + \eta_i^f [z_i(x) - u_{n-1}(x)], \quad v_i^f(s) = \frac{32 - z_i(s)}{R_i(s)}$$

The complete temperature profile in the layered soil is given by

$$u_n(x) = U_i(x) v_i^w(x) + w_i(x), \quad x \in [x_{i-1}, x_i], \quad x < s$$

$$u_n(x) = R_i(x) v_i^f(x) + z_i(x), \quad x \in [x_{i-1}, x_i], \quad x > s.$$

If $\alpha(n\Delta t) < 32$ then no water phase is present. In this case we can dispense with the forward sweep involving U_i and w_i and with the functional (3.29); instead we carry out the forward sweep for the ice phase by computing R_i and z_i. At the surface $x = 0$ the temperature is $u^f(0) = \alpha(n\Delta t)$. Hence in view of the Riccati transformation (3.28) the gradient $v_1^f(0)$ must me chosen such that

$$\alpha(n\Delta t) = R_1(0) v_1^f(0) + z_1(0),$$

or

$$v_1^f(0) = \frac{\alpha(n\Delta t) - z_1(0)}{R_1(0)} .$$

This value for v_1^f and $s = 0$ are used to complete the backward sweep by integrating (3.33).

In summary, we have a well defined (analytical) solution algorithm for the free interface problem (2.1) of heat transfer with change of phase in a layered soil. It remains to show that this algorithm, which was presented only formally, is actually applicable. Thus we need to show that given the solution $\{u_{n-1}(x), s_{n-1}\}$ at the previous time level we can actually compute $\{u_n, s_n\}$ with this algorithm.

In order to demonstrate that the above sweep method can be used to compute a solution $\{u_n, s_n\}$ three properties of the above equations must be established, namely that the Riccati equations (3.26) and (3.29) have bounded solutions, and that the functional equation (3.31) has a root. If this is indeed the case, then all other equations have solutions since they are linear ordinary differential equations. The existence question for the Riccati equations is readily resolved. It is seen from (3.26) that $U_1(0) = 0$ and $U_1'(0) = 1$ and $U_1'(\hat{x}) = 1$ whenever $U_1(\hat{x}) = 0$. Hence $U_1 > 0$ on $[0, x_1]$. Moreover $U_1' < 0$ if $U > \sqrt{\frac{1}{\eta_1^w}}$ which assumes that $U_1 \leq \frac{1}{\sqrt{\eta_1^w}}$ on $[0, x_1]$. A similar analysis can be applied to U_2 and successively to U_j, $j = 3, \ldots, M$. A similar argument shows that $-\infty < R_j < 0$ on $[0, X]$.

The next problem to be tackled is the existence of a root s for the functional (3.29). Since this function does not come into play if the surface temperature is below freezing let us suppose that $\alpha(n\Delta t) > 32$. It follows from (3.29) and (3.31) that $\lim_{x \to 0+} U_1(x) = 0+$ and $\lim_{x \to 0+} \phi(x) = -\infty$, and similarly that $\lim_{x \to X-} \phi(x) = +\infty$, regardless of where s_{n-1} may be located. The question thus would resolve itself if ϕ were continuous. Because z_j and w_j are continuous at fixed interfaces and $k_j^w U_{j+1}(x_j) = k_{j+1}^w U_j(x_j)$ and $k_j^f R_{j-1}(x_{j-1}) = k_{j-1}^f R_j(x_{j-1})$ we see that the first two terms of (3.31) are continuous. The last term, however, exhibits jumps at the interfaces because in general $\lambda_j \rho_j \neq \lambda_{j+1} \rho_{j+1}$. From a conceptual point of view this behavior does not cause any difficulties. We shall simple trace $\phi(x)$ from $x = 0$ until it first turns nonnegative; this point is well defined if ϕ is taken to be continuous from the left at the interfaces.

In summary, then, the above algorithm constitutes a feasible mathematical method to compute the solution $\{u_n, s_n\}$ for the free interface problem (2.4) since at each time level all equations have solutions.

Two important aspects of this problem remain to be investigated. The first of these is of overriding concern in an industrial environment. How does one obtain actual (i.e., numerical) answers from the above method? After all, the method is stated in terms of initial value problems for ordinary differential equations. Hence methods for their integration must be considered. In addition, it is necessary to find the zero crossing of a nonlinear, possibly discontinuous function. This problem also needs some comments. We shall next discuss in some detail how to obtain numerical answers with the above method. The second aspect concerns convergence of the computed answers to the solution of the continuous problem. This point is generally of interest to the mathematician and will be discussed in the last section; non-mathematicians usually decide on convergence on the basis of the computed results as illustrated below.

The numerical solution of the forward and backward sweep equations can be obtained quite simply. Since in replacing the continuous Stefan problem with the "by lines approximation" a simple backward difference quotient was employed, the truncation error of the approximation is $0(\Delta t)$. Hence it seems more than adequate to apply a second order integrator for the spacial variable. Moreover, it is generally observed that implicit integration techniques exhibit better stability properties than explicit methods. The drawback of implicit methods is due to the fact that an equation must be solved for the unknown. If this equation is nonlinear this problem can be complicated. Fortunately, this difficulty does not occur if the trapezoidal rule is applied to carry out the sweeps. Let us trace out the application of the trapezoidal rule for the equations at hand. For definiteness, let us suppose we need to find $U_1(x)$ and $w_1(x)$ over $[0,x_1]$. We define a partition $\{0 = x^o < x^1 < \dots < x^m = x_1\}$ of $[0,x_1]$ and set $\Delta x_j = x^j - x^{j-1}$, $j = 1,\dots,M$. Then U_1 and w_1 at the mesh points $\{x^j\}$ are found recursively from the trapezoidal formula

$$U_1(x^j) = U_1(x^{j-1}) + \frac{\Delta x_j}{2}[2-\eta_1{}^w(U_1{}^2(x^j) + U_1{}^2(x^{j-1}))], \quad U_1(x^o) = 0$$

$$\begin{aligned} w_1(x^j) = w_1(x^{j-1}) + \frac{\Delta x_j}{2}[&-\eta_1{}^w U_1(x^j)(w_1(x^j)-u_{n-1}(x^j)) \\ &-\eta_1{}^w U_1(x^{j-1})(w_1(x^{j-1})-u_{n-1}(x^{j-1}))], \quad w_1(x^o) = \alpha(n\Delta t). \end{aligned}$$

The first of these equations is quadratic and can be solved in closed form for the unknown $U_1(x^j)$; once $U_1(x^j)$ is known the second equation is linear in $w_1(x^j)$ and can be solved without difficulty. Similar expressions are used for R_j and z_j. In this manner, U_i, w_i, R_j and z_j can be computed at discrete mesh points distributed over the total interval $[0,X]$.

Once U_i, R_j, w_i and z_j are known at the mesh points over the entire interval $[0,X]$ the value of Φ is known at the mesh points. In order to find the free interface where $\Phi(x) = 0$ we may simply evaluate the functional at each mesh point and place s by linear interpolation between the first two mesh points $\{x^k, x^{k+1}\}$ for which $\Phi(x^k) \cdot \Phi(x^{k+1}) \leqq 0$. In general, s will not coincide with a mesh point of the partition. In this case several options exist. Either s is added to the partition and $U_i(s)$, $w_i(s)$, $R_j(s)$, and $z_j(s)$ are computed with the trapezoidal rule, or s is moved to the nearest mesh point. The latter course is quite simple and proved adequate. Finally, once the location of s is fixed the trapezoidal rule is used to integrate v_i^w and v_j^f which then are used to define the complete temperature profile $u_n(x)$ over $[0,X]$ at time $t = n\Delta t$.

Comment: The forward/backward sweep method introduced above is commonly called the method of invariant imbedding. A detailed discussion of this initial value solution algorithm may be found in G. H. Meyer, Initial Value Methods for Boundary Value Problems, Academic Press, 1973.

Suggested exercises:

1) Verify by back-substitution the correctness of the invariant imbedding equations for an interface problem.

2) Find the closed form solution of the Riccati equation (3.26).

3) Find the asymptote to the solution of $U' = 1 + 4U - U^2$, $U(0) = 0$.

4) Discuss the difficulties in applying the trapezoidal rule to $u' = \sin(u + t)$, $u(0) = 0$.

5) Set up the invariant imbedding equation for the problem $u_{xx} - u_t = 0$, $u(x,0) = u_0(x)$, $\frac{\partial u}{\partial x}(0,t) = \alpha(t)$, $u(1,t) = g(t)$.

6) Write out the invariant imbedding equations for the free boundary value problem of Section 2, exercise 3.

4. An Application

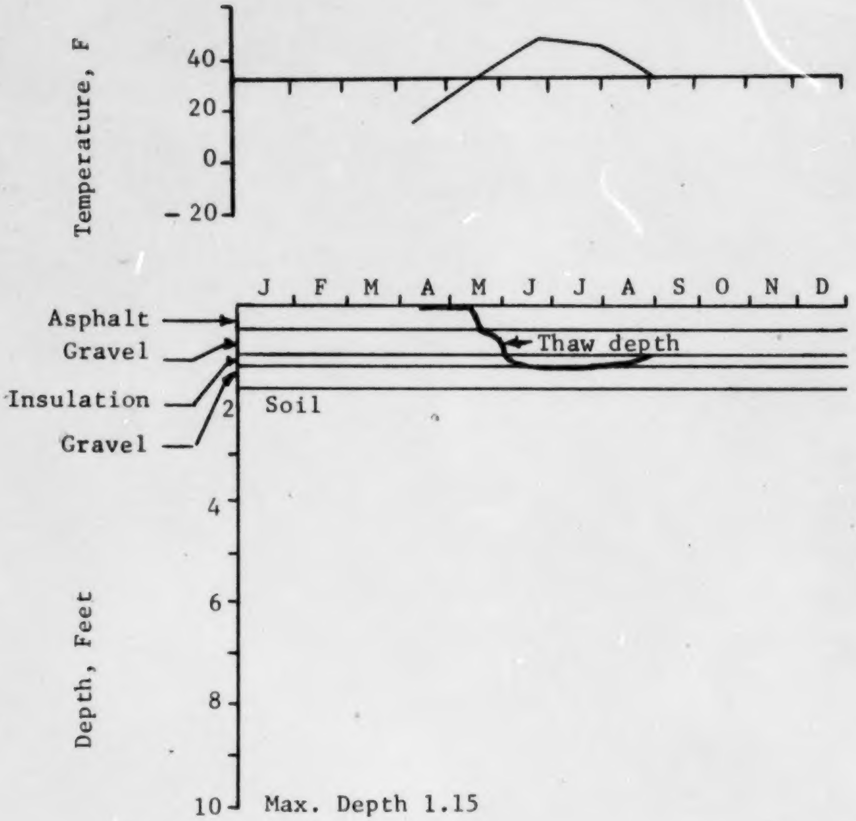
The above model and numerical solution technique were used for the evaluation of various design alternatives for an airstrip on permafrost. The model was considered applicable because thermal edge effects do not appear to extend beyond four to five feet from the edge of the runway so that heat flow under the load bearing portion of the structure is essentially one dimensional.

Because there is some uncertainty in the data necessary to carry out the computation, all parameters had to vary over considerable ranges which necessitated a lot of computation. Fortunately, the facilities available and the speed of the computer algorithm allowed the effective use of computer graphics. Rather than on reams of paper the computed results were displayed immediately after computation on a television screen. Two illustrations, redrawn from photographs of the television screen show representative results for a structure of four man-made layers on a permafrost soil for two different temperature inputs.

Extensive experiments had shown that the results are seasonally periodic and that no long range effects accumulate over large time spans (up to 40 years). Therefore, only one cycle is exhibited. The computation starts in April and continues until the surface temperature reaches the freezing point. The main quantity of interest is, of course, the maximum thaw depth; since it always had begun decreasing at this time the computation was terminated at this point (between the early and middle part of September).

Figure 2 shows a plot of the surface temperature $\alpha(t)$ vs. time (the time axis is horizontal and labeled with J(anuary), F(ebruary), etc.) and a plot of the computed thaw depth $s(t)$ through a structure of asphalt, gravel, an insulator-like styrofoam, and gravel resting on a soil in a state of permafrost. As is seen, for the thermal parameters displayed on the right of the illustration the interface $s(t)$ does not enter the soil.

Figure 3 shows the same system when the summer temperatures are increased by 10°F. The interface is now seen to enter the moisture bearing gravel underneath the insulator; however, the permafrost still escapes the summer thaw. Because of the speed of the algorithm many other temperature inputs $\alpha(t)$ were tried. This was important because of the uncertainty of the soil surface temperature. It has been found that the annual median soil temperature is

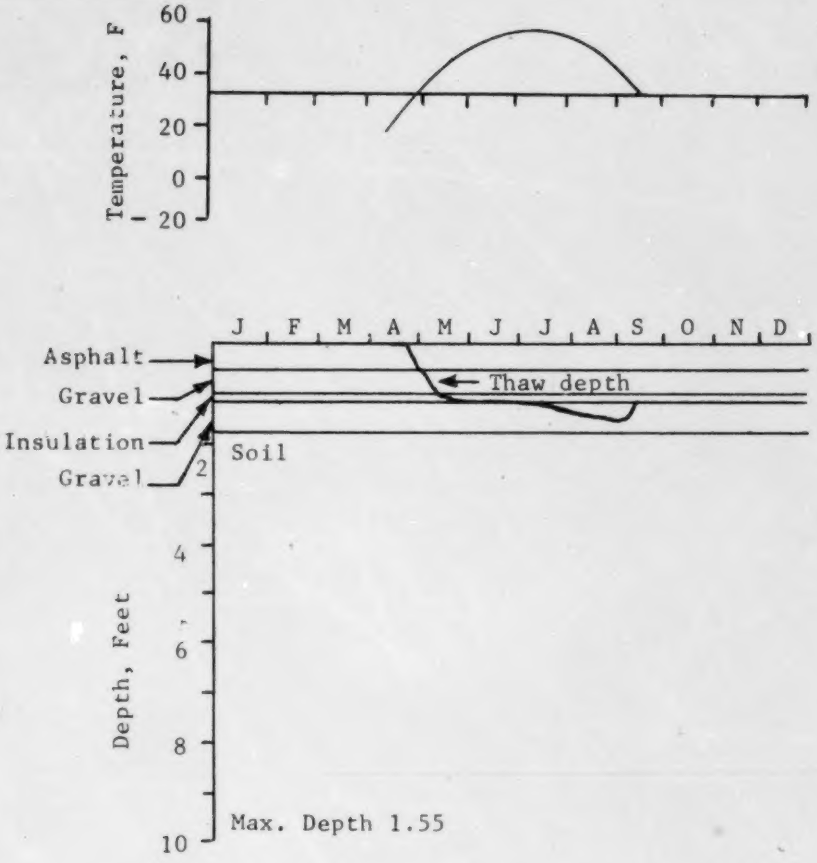


LYR	MATERIAL	THICK	DEPTH
1	Asphalt	.5	.5
2	Gravel	.5	1.0
3	Insulation	.2	1.2
4	Gravel	.5	1.7
5	Soil	200.0	201.7

LYR	CONDUCTIVITY ICE	CONDUCTIVITY WATER	SAT.	POR.
1	20.00	20.00	0.00	.20
2	75.00	50.00	1.00	.35
3	1.20	1.20	0.00	.20
4	75.00	50.00	1.00	.35
5.	60.00	40.00	1.00	.50

MONTH	TEMPERATURE
JAN	-9.9
FEB	-13.3
MAR	-7.7
APR	8.6
MAY	25.9
JUN	39.1
JUL	47.5
AUG	45.2
SEP	36.0
OCT	22.6
NOV	6.5
DEC	-3.7

Figure 2. Thaw Depth Under Asphalt-Topped Gravel-Insulation Sandwich



LYR	MATERIAL	THICK	DEPTH
1	Asphalt	.5	.5
2	Gravel	.5	1.0
3	Insulation	.2	1.2
4	Gravel	.5	1.7
5	Soil	200.0	201.7

	CONDUCTIVITY			
LYR	ICE	WATER	SAT.	POR.
1	20.00	20.00	0.00	.20
2	75.00	50.00	1.00	.35
3	1.20	1.20	0.00	.20
4	75.00	50.00	1.00	.35
5	60.00	40.00	1.00	.50

MONTH	TEMPERATURE
JAN	-9.9
FEB	-13.3
MAR	-7.7
APR	8.6
MAY	36.0
JUN	50.0
JUL	57.5
AUG	55.0
SEP	46.0
NOV	6.5
DEC	-3.7

Figure 3. Thaw Depth Under Asphalt-Topped Gravel-Insulation Sandwich with Abnormally High Surface Temperature

several degrees higher than the corresponding air ground temperature. However, daily records are needed for this model which are not yet known. It also has been attempted to include convective Newtonian cooling on the surface in the form of

$$k\frac{\partial u}{\partial x}(0,t) = h(t)(u(0,t) - \alpha(t))$$

where the heat transfer coefficient h is adjusted according to snow cover and wind velocity. Again reliable data are lacking, so that the computed results are taken to indicate trends rather than absolute values.

In carrying out these computations the influence of the time and space steps on the computed answers was examined. It is generally believed than an implicit approximation of the type discussed here is stable; hence it is common to choose the mesh sizes sufficiently small so that the computer answers change little on further mesh refinements, and sufficiently large so that the computation is rapid. Practical experience will usually point to the optimum mesh sizes.

Suggested exercises:

1) Discuss how the above program can be incorporated into a complete model for the most economical construction of an airstrip on permafrost.

2) Integrate $u_{xx} - u_t = 0$, $u(0,t) = t$, $u(1,t) = 1$, $u(x,0) = x$ numerically with

   a) a forward finite difference scheme

   b) the method of lines

   In both cases examine the behavior of the computed solutions as $\Delta t, \Delta x \to 0$.

3) Modify the model and solution technique for a layered medium in order to include convective surface cooling of the form

$$k\frac{\partial u}{\partial x}(0,t) = h(t)(u(0,t) - \alpha(t)).$$

4) formulate the Stefan problem for the freezing of a large lake, write the computer program and examine the effect of the various thermal and numerical parameters.

Comments:

The reader interested in further details on the parameter study is referred to the account of G. H. Meyer, N. N. Keller, and E. J. Couch, Thermal Model for Roads, Airstrips, and Building Foundations in Permafrost Regions, Journal of Canadian Petroleum Technology, April-June 1972, and to the references given there.

Additional data on soil and permafrost may be found in the volume on permafrost listed at the end of Section 1 and in the monograph of M. S. Kersten, Thermal Properties of Soils, Bull. 28, Eng. Exp. Sta., University of Minnesota, 1949.

Convergence: From an engineering and management point of view the model developed above represented a fairly complete description of heat transfer in a layered soil. It was demonstrated that this model has a solution $\{u_n, s_n\}$ at each time level and that the solution can be obtained numerically by quite elementary techniques. For a mathematician the problem does not end here since no assurance has been given that the approximate solutions do indeed come close, in some sense, to the continuous solution of the Stefan problem (2.9). It is still required to prove convergence of $\{u_n, s_n\}$ to $\{u(x,n\Delta t), s(n\Delta t)\}$ as $\Delta t \to 0$. What is gained from such a convergence proof is an understanding of why the solution method works, as well as an assurance that the numerical results converge to the true solution and not to some spurious functions. It may also indicate that applied mathematics, even in an industrial environment, can go considerably beyond calculus.

With our present understanding this convergence proof cannot be carried through without imposing several hypotheses on the model which may not necessarily be justified by the physics of heat transfer with change of phase and, in fact, may not be satisfied by our model. Thus the convergence proof given here must be taken only as an indication that the method of lines approximation can be made to work at least in certain special cases.

In order to keep the problem tractable we shall make the simplification that the model consists of only one layer with prescribed surface temperature $\alpha(t)$ and bottom temperature U_X. Let us index all thermal quantities above the interface $s(t)$ with 1, those below $s(t)$ by 2. Then the continuous and by-lines equations at the nth time level are

$$k_i u_{i_{xx}} - c_i \rho_i u_{i_t} = 0; \quad k_i u_i'' - \frac{c_i \rho_i}{\Delta t}[u_i - u_{n-1}(x)] = 0 \quad i = 1, 2$$

$$u_1(t) = \alpha(t), \ u_2(X) = U_X; \quad u_1(0) = \alpha(n\Delta t), \ u_2(X) = U_X$$

$$u(x_1 0) = u_o(x), \ u(s(0),0) = 0$$

with the free interface condition

$$0 = k_1 u_{1x} - k_2 u_{2x} + \rho_1 \lambda \frac{ds}{dt}; \quad k_1 u_1' - k_2 u_2' + \rho_1 \lambda \frac{s - s_{n-1}}{\Delta t} = 0$$

$$u_1(s(t),t) = u_2(s(t),t) = u_1(s) = u_2(s) = 0.$$

(For ease of exposition we have chosen $u = 0$ as the phase transition temperature.) It is known that this problem has a unique classical solution.

Let us state the final convergence theorem at this time.

Theorem: If $u_1(x,t)$, $u_2(x,t)$ and $s(t)$ are the solutions of the above continuous free interface problem, and if $U_1^N(x,t)$, $U_2^N(x,t)$, $s^N(t)$ are approximate solutions defined by setting

$$U_i^N(x,t) = u_i(x),$$

$$s^N(t) = \frac{t_n - t}{\Delta t} s_{n-1} + \frac{t - t_{n-1}}{\Delta t} s_n \quad t \in (t_{n-1}, t_n)$$

then $U_1^N(x,t) \to u_1(x,t)$, $U_2^N(x,t) \to u_2(x,t)$ and $s^N(t) \to s(t)$ (in a weak sense) as $\Delta t = \frac{T}{N} \to 0$ for some fixed final time T.

Considerable ground work and a certain sophistication are required for the proof of this theorem. It will be carried out in detail in the appendix.

Appendix

Convergence proof for the method of lines solution for the Stefan problem.

If $u_1(0) > 0$, $u_2(X) < 0$ then the previous section shows that $\{u_1, u_2, s\}$ exists at each time level. Convergence will now be proved under the following hypotheses:

1) $s(0) \in (\epsilon, X-\epsilon)$ for some $\epsilon > 0$

2) $u_o(x)$ is strictly monotone decreasing

3) If we set $\hat{K} = \max\{\max\limits_{t\in[0,T]} \alpha(t), -u_X\}$,

$$\text{then} \quad \hat{K}\,\frac{k_2 c_1}{k_1 \lambda} < 1 \quad \text{and} \quad \hat{K}\,\frac{c_2 \rho_2 k_1}{k_2 \rho_1 \lambda} < 1$$

The third of these hypotheses has no reasonable physical interpretation and is imposed only to make the following proof work. It may be noted, however, that for a moisture bearing soil the inequalities are unually satisfied. Indeed, it generally is true that $|\frac{k_i}{k_j}| \leq 2$, $|\frac{\rho_i}{\rho_j}| \approx 1$ and $c_1 < 1$ for $i, j, = 1, 2$ if cgs units are used. Since $\lambda = 80$ it follows that $\alpha(t)$ and u_X may vary as much as 40°C from the freezing temperature of 0°C. Arctic summer temperatures, i.e., $\alpha(t)$, do not reach such high values while u_X is held fixed above -40°C. Thus the hypotheses are reasonable for the summer melting of a soil.

We can now give the key lemma for our convergence proof.

Lemma: Suppose that the above hypotheses apply. Let η be chosen so large that $|\alpha(t)| \leq \hat{K}(1-e^{-\eta\epsilon})$, $|u_i^o(x)| \leq \hat{K}(1-e^{-\eta|x-s_o|})$. Then $0 \geq u_i^{n'}(s_n) \geq -\eta\hat{K}$ for $n = 1, \ldots, N_o$, where N_o is the largest integer such that $s_n \in [\epsilon, X-\epsilon]$ for $n = 1, \ldots, N_o$, and where u_i^n denotes u_i at temperature t_n.

Proof. We shall prove by induction that $|u_i^n(x)| \leq \hat{K}(i-e^{-\eta|x-s_n|})$ which will allow us to bound $u_i'(s_n)$. If $n = 0$ this inequality holds by hypothesis. Suppose that it is true for an arbitrary integer n-1.

Consider then the functions

$$W_1(x) = u_1^n(x) - \hat{K}(1-e^{-\eta(s_n-x)})$$

$$W_2(x) = u_2^n(x) + \hat{K}(1-e^{-\eta(x-s_n)})$$

If we define the operator $L_i\Phi = \Phi'' - \frac{c_i\rho_i}{k_i\Delta t}\Phi$, then it follows from the above equations that

$$L_1W_1 = \frac{c_1\rho_1}{k_1\Delta t} u_1^{n-1}(x) + \hat{K}\eta^2 e^{-\eta(s_n-x)} + \frac{c_1\rho_1}{k_1\Delta t}\hat{K}(1-e^{-\eta(s_n-x)})$$

$$L_2W_2 = -\frac{c_2\rho_2}{k_2\Delta t} u_2^{n-1}(x) - \hat{K}\eta^2 e^{-\eta(x-s_n)} + \frac{c_2\rho_2}{k_2\Delta t}\hat{K}(1-e^{-\eta(x-s_n)})$$

By hypothesis $u_1^{n-1}(x) \le \hat{K}(1-e^{-\eta(s_{n-1}-x)}) =$

$$\hat{K}(1-e^{-\eta(s_n-x)+\eta(s_n-s_{n-1})}),$$

$$u_2^{u-1}(x) \ge \hat{K}(1-e^{-\eta(x-s_{n-1})}) = \hat{K}(1-e^{-\eta(x-s_n)-\eta(s_n-s_{n-1})})$$

so that

$$L_1W_1 \ge \hat{K}e^{-\eta(s_n-x)}[\eta^2 + (e^{\eta(s_n-s_{n-1})}-1)\frac{c_1\rho_1}{k_1\Delta t}$$

$$L_2W_2 \le \hat{K}e^{-\eta(x-s_n)}[-\eta^2 + \frac{c_2\rho_2}{k_2\Delta t}(1-e^{-\eta(s_n-s_{n-1})})].$$

For definiteness let us assume that $s_n \le s_{n-1}$, then $L_2W_2 \le 0$; moreover, by inspection we see that $W_2(s_n) = 0$, $W_2(X) \ge 0$. The maximum principle applied to L_2W_2 shows that W_2 cannot have a negative minimum on (s_n,X) since at such a minimum L_2W_2 would be positive. Hence $W_2 \ge 0$. In particular, this implies that $W_2'(s_n) = u_2^{n'}(s_n) + \hat{K}\eta \ge 0$ or $-\hat{K}\eta \le u_2^{n'}(s_n)$. Since $u_1^{n'}(s_n) \le 0$ it follows from

$$-k_1u_1^{n'}(s_n) + k_2u_2^{n'}(s_n) = \lambda\rho\frac{s_n-s_{n-1}}{\Delta t}$$

that

$$s_n-s_{n-1} \le \frac{\Delta t}{\lambda\rho}k_2\eta\hat{K}$$

If this expression is substituted into the estimate for L_1W_1 we obtain

$$L_1W_1 \ge \hat{K}e^{-\eta(s_n-x)}[\eta^2 + \frac{c_1\rho_1}{k_1\Delta t}(\exp(\eta(\frac{-\Delta t k_2\eta\hat{K}}{\lambda\rho})-1) \ge$$

$$\hat{K}e^{-\eta(s_n-x)}[\eta^2 - \frac{c_1\rho_1}{k_1}\frac{\Delta t}{\Delta t}\frac{k_2\eta^2\hat{K}}{\lambda\rho}]$$

or

$$L_1W_1 \ge \hat{K}\eta^2 e^{-\eta(s_n-x)}[1 - \frac{c_1\rho_1 k_2}{k_1\lambda\rho}\hat{K}]$$

By hypothesis 3 it follows that $L_1W_1 \geqq 0$; the maximum principle together with $W_1(0) \leqq 0$, $W_1(s_n) = 0$ implies that W_1 has no positive maximum on $(0,s_n)$; hence $u_2^n \leqq \hat{K}(1-e^{-\eta(s_n-x)})$ and $u_1^{n'}(s_n) \geqq -\hat{K}\eta$.

Corollary: $|s_n - s_{n-1}| \leqq (\hat{K}\eta/_{\lambda\rho})\ \Delta t$

Proof. If $s_n \leqq s_{n-1}$ the proof of the above lemma contains this inequality. If $s_n \geqq s_{n-1}$ the roles of W_1 and W_2 in the proof of the preceding lemma are reversed; we observe that $L_1W_1 \geqq 0$ and compute that $L_2W_2 \leqq 0$. In both cases we find that $|u_i^{n'}(s_n)| \leqq \hat{K}\eta$ so that $|s_n - s_{n-1}| \leqq \frac{\hat{K}\eta}{\lambda\rho}\,\Delta t$.

The preceding estimate allows us to establish the existence of a Lipschitz continuous free interface $s(t)$. For given $N > 0$ and increment $\Delta t = \frac{T}{N}$ let us define the polygonal path $S_N(t)$ as the piecewise linear continuous function obtained by connecting two adjacent points with a straight line. We note that for arbitrary $n = N$, S_N is given by

$$S_N(t) = \frac{1}{\Delta t}\{(t-t_n)s_{n+1} + (t_{n+1}-t)s_n\}, \quad t \in [t_n, t_{n+1}].$$

Let us look at the collection $\{S_N(t)\}$ for $N = 1,2,\ \ldots$ of all piecewise linear paths. Since $S_N(t) \in (\varepsilon, X-\varepsilon)$ for each N, all these functions are uniformly bounded. Moreover, each S_N is piecewise differentiable and $|S_N'(t)| \leqq \frac{\hat{K}\eta}{\rho\lambda}$ except possibly at the corner points $\{t_n\}$ where S_N is continuous. Hence

$$|S_N(t) - S_N(s)| \leqq \frac{\hat{K}\eta}{\rho\lambda}(t-s)$$

so that $\{S_N\}$ is a family of equicontinuous, uniformly bounded functions defined on $[0,T]$. By the Ascoli-Arzela theorem there exists a continuous function $s(t)$ and a subsequence $\{S_{N_e}(t)\}$ such that $S_{N_e} \to s$ uniformly on $[0,T]$ as $N_e \to \infty$. It needs to be shown that s is actually the free interface.

In what follows, whenever we talk about convergence we shall assume that $\Delta t_{N_e} = \frac{T}{N_e}$ so that $\Delta t_{N_e} \to 0$ (subsequently always written as $\Delta t \to 0$) implies that $S_{N_e}(t) = S_N(t) \to s(t)$ uniformly. In other words, we are always talking about convergence of a subsequence of $\{u_n, s_n\}$ of the by-lines aporoximation.

It may be noted that the classical formulation of the Stefan problem contains the derivative of $s(t)$ while the interface found so far is merely Lipschitz continuous. In order to verify that s is actually the free interface of the Stefan problem one must either establish that $s(t)$ is differentiable and satisfies the given boundary conditions, or one must relax the concept of a solution for the Stefan problem so that $s'(t)$ does not occur. We shall choose the second course by defining a "weak solution" for the interface problem (2.1).

The motivation for the definition derives from the following observation. Suppose that $u(x,t)$ is a twice continuously differentiable solution of the classical Stefan problem, and let $s(t)$ denote the free interface. Let $\Phi(x,t)$ be any twice continuously differentiable function on $[0,X] \times [0,T]$ which vanishes for $x = 0$, $x = X$ and $t = T$. Φ will subsequently be called a test function. Since $k_1 u_{xx} - c_1\rho_1 u_t = 0$ on $(0,s(t))$, $t \in (0,T)$ it follows from integrating by parts and use of the boundary conditions imposed on u and ρ that

$$0 = \int_0^T\int_0^{s(r)} \Phi(x,r)[k_1 u_{xx} - c_1\rho_1 u_t]\,dx\,dr = k_1\int_0^T\int_0^{s(r)} u(x,r)\Phi_{xx}(x,r)\,dx\,dr +$$

$$k_1\int_0^T \Phi(s(r),r)u_x(s(r),r)\,dr + k_1\int_0^T \alpha(r)\Phi_x(0,r)\,dr +$$

$$c_1\rho_1 \int_0^T\int_0^{s(t)} \Phi_r(x,t)u(x,t)\,dx\,dr + c_1\rho_1 \int_0^{s(0)} \Phi(x,0)u_0(x)\,dx.$$

Integrating $\int_0^T\int_{s(r)}^X k_2\Phi(x,r)[k_2 u_{xx} - c_2\rho_2 u_r]\,dx\,dr = 0$ by parts and adding the resulting expression to the preceding line we obtain

$$(4.1) \qquad H(u,\Phi,s) \equiv \int_0^T\int_0^{s(r)} [k_1\Phi_{xx}(x,r) + c_1\rho_1\Phi_r(x,r)]u(x,r)\,dx\,dr +$$

$$\int_0^T\int_{s(r)}^X [k_2\Phi_{xx}(x,r) + c_2\rho_2\Phi_r(x,r)]u(x,r)\,dx\,dr - k_1 \int_0^T \alpha(r)\Phi_x(0,r)\,dr +$$

$$k_2 u_X \int_0^T \Phi_x(X,r)\,dr + c_1\rho_1 \int_0^{s(0)} \Phi(x,0)u_0(x)\,dx + \int_{s(0)}^X c_2\rho_2\Phi(x,0)u_0(x)\,dx -$$

$$\lambda\rho \int_0^T \Phi(s(r),r)s'(r)\,dr = 0.$$

It may be noted that (4.1) does not involve any derivatives of u; moreover, even though $s'(r)$ is present, $s'(r)$ need merely be a Riemann integrable function and not necessarily continuous as in the classical formulation.

We shall now reverse the process of linking (4.1) with (2.1)

Definition. A pair $\{u(x,t),s(t)\}$ of bounded measurable functions which satisfies (4.1) for all twice continuously differentiable Φ defined on $[0,X] \times [0,T]$ and vanishing for $x = 0$, $x = X$ and $t = T$ is called a weak solution of the Stefan problem (2.1).

It is not difficult to show that if $\{u(x,t),s(t)\}$ is a smooth weak solution then it necessarily must satisfy the classical Stefan problem. For example, suppose that $u(x,t)$ is twice continuously differentiable in x and continuously differentiable in t. Let $\Phi(x,t)$ be any smooth function with compact support on the set $0 = \{(x,t) : x \in (0,s(t)),\ t \in (0,T)\}$. Substitution into (4.1) and integration by parts yields

$$\int_o^T\int_o^{s(r)} (k_1u_{xx} - c_1\rho_1u_t)\Phi(x,r)dxdr = 0$$

since all other terms vanish due to the limited support of $\Phi(x,t)$. If $k_1u_{xx}-c_1\rho_1u_t \neq 0$ at some point (x_o,t_o) of D then we can find a Φ whose support is contained in a neighborhood of (x_o,t_o) where $k_1u_{xx}-c_1\rho_1u_t \neq 0$ which would yield a positive value of the above equation. Since this is not permissible we conclude that $k_1u_{xx}-c_1\rho_1u_t = 0$ on D. Similarly, one shows that $k_2u_{xx}-c_2\rho_2u_t = 0$ is satisfied to the right of $s(t)$. To demonstrate that the boundary values are achieved one again integrates by parts, collects integrals over common domain and then chooses Φ such that all integrals except the one over the boundary in question vanish.

Convergence of $\{u_n,s_n\}$ to a weak solution of the Stefan problem can now be established. For given N let us define the function

$$U_N(x,t) = u_n(x), \quad t \in (t_{n-1},t_n), \quad n = 1, \ldots, N.$$

It is seen that the sequence $\{U_N\}$ is a uniformly bounded sequence of functions in the Hilbert space L_2 of square integrable functions on $[0,X] \times [0,T]$. Because of its boundedness it contains a weakly convergent subsequence $\{U_{N_k}\}$, i.e.,

$$\lim_{N_k\to\infty} \int_o^T\int_o^X \Phi(x,t)U_{N_k}(x,t)dxdt = \int_o^T\int_o^X \Phi(x,t)u(x,t)dxdt$$

for some $u \in L_2$ and all $\Phi \in L_2$. In what follows we shall deal only with the subsequence $\{N_k\}$ and omit the subscript k. Notice that $\{N_k\}$ is actually a subsequence of the sequence $\{N_e\}$ which led to a convergent sequence $\{S_{N_e}(t)\}$.

If for given N (or equivalently, Δt) and arbitrary test function Φ the approximation

$$\Phi_N(x,t) = \frac{1}{\Delta t}\left[\Phi(x,t_n)(t_{n+1}-t) + \Phi(x,t_{n+1})(t-t_n)\right],$$

$$t \in [t_n, t_{n+1})$$

as well as Φ_N, U_N and S_N are substituted into (4.1) for Φ, U, and S, then it can be observed that the limit of (4.1) exists as $N \to \infty$. For example, consider the first integral of (4.1). Let

$$J = \lim \int_0^T \int_0^{S_N(r)} U_N\left[k_1 \Phi_{N_{xx}} + c_1 \rho_1 \Phi_{N_r}\right]dx\,dr -$$

$$\int_0^T \int_0^{s(r)} \left[k_1 \Phi_{xx} + c_1 \rho_1 \Phi_r\right] u(x,r)\,dx$$

If we define $L_i^* \Phi = k_i \Phi_{xx} + c_i \rho_i \Phi_t$ the above expression can be rewritten as

$$J = \lim_{N\to\infty} \Big[\int_0^T \Big\{\int_0^{S_N(r)} U_N L_1^* \Phi_N dx - \int_0^{s(r)} U_N L_1^* \Phi_N dx +$$

$$\int_0^{s(r)} U_N L_1^* \Phi_N dx - \int_0^{s(r)} U_N L^*{}_1 \Phi dx +$$

$$\int_0^{s(r)} U_N L_1^* \Phi dx - \int_0^{s(r)} U L_1^* \Phi dx\Big]dr$$

$$J = \lim \int_0^T \Big\{\int_{s(r)}^{S_N(r)} U_N L_1 \Phi_N dx + \int_0^{s(r)} U_N (L_1 \Phi_N - L_1 \Phi) dx +$$

$$\int_0^{s(r)} (U_N - u) L_1 \Phi dx\Big\}dr$$

The first integral vanishes because U_N and $L^* \Phi_N$ are bounded and $S_N \to s$ uniformly in t. The second integral vanishes because $L^* \Phi_N \to L^* \Phi$ as $\Delta t \to 0$, and the last integral vanishes because of the weak convergence of U_N. Hence $J = 0$; the other integrals of $H(U_N, \Phi_N, S_N)$ are handled analogously, and as a consequence

$$H(U_N,\Phi_N,S_N) \rightarrow H(u,\Phi,s) \quad \text{as} \quad N \rightarrow \infty$$

It must be shown that $H(u,\Phi,s) = 0$. Let us approximate $H(U_N,\Phi_N,S_N)$ by a Riemann sum (with respect to time), i.e., we shall write

$$H_N(U_N,\Phi_N,S_N) = \sum_{n=1}^{N} \Delta t\{\int_o^{s_n} U_n(L_1^*\Phi_N)(x,t_n) +$$

$$\int_{s_n}^{X} U_n(L_2^*\Phi_N)(x,t_n) - k_1\alpha(t_n)\Phi_{N_x}(0,t_n) + k_2U_X\Phi_{N_x}(X,t_n) -$$

$$\lambda\rho\Phi_N(s_n,t_n)\,\frac{s_n-s_{n-1}}{\Delta t}\}$$

$$+\, c_1\rho_1\int_o^{s_o}\Phi(x,0)U_o(x)dx + c_2\rho_2\int_{s_o}^{X}\Phi(x,0)U_o(x)dx,$$

where $L_i^*\Phi_N(x,t_n) = k_i\Phi_{xx}(x,t_n) + c_i\rho_i\,\dfrac{\Phi(x,t_{n+1})-\Phi(x,t_n)}{\Delta t}$.

It can be verified that $H_N(U_N,\Phi_N,S_N) \rightarrow H(U_N,\Phi_N,S_N)$ as $N \rightarrow \infty$. For example,

$$\Big|\sum_{n=1}^{N}\Delta t\; k_1\int_o^{s_n} U_n\Phi_{xx}(x,t_n)dx - \int_o^T\int_o^{s(r)} U_Nk_1\Phi_{xx}(x,r)dr\Big| =$$

$$\sum_{n=1}^{N} k_1\{\Delta t\int_o^{s_n} U_n\Phi_{xx}(x,t_n)dx - \int_{t_{n-1}}^{t_n}[\int_o^{s_n} U_n\Phi_{xx}(x,r)dx +$$

$$\int_{s_n}^{s(r)} U_n\Phi_{xx}(x,r)dx]dr$$

$$\leq k\Delta t \quad \text{because} \quad \Phi_{xx}(x,t_n) -$$

$$\Phi_{xx}(x,r) \leq k\Delta t \quad \text{and} \quad |s_n-s(r)| \leq k\Delta t.$$

The other terms are handled analogously. Hence it is established so far that $H_N(U_N,\Phi_N,S_N) \rightarrow H(U_N,\Phi_N,S_N) \rightarrow H(u,\Phi,s)$. Finally, integrating by parts the following expressions and using the boundary and interface conditions for U_n leads to

$$0 = \int_o^{s_n} [k_1 u_n'' - \frac{c_1\rho_1}{\Delta t} (u_n - u_{n-1})]\Phi(x,t_n)dx +$$

$$\int_o^{s_n} [k_2 u_n'' - \frac{c_2\rho_2}{\Delta t} (u_n - u_{n-1})]\Phi(x,t_n)dx =$$

$$\int_o^{s_n} k_1 u_n \Phi_{xx}(x,t_n)dx + \int_o^{s_n} k_2 u_n \Phi_{xx}(x,t_n)dx +$$

$$k_2 u_X \Phi_x(X,t_n) - k_1 \alpha(t_n)\Phi_x(0,t_n)$$

$$- \int_o^{s_n} \frac{c_1\rho_1}{\Delta t} (u_n - u_{n-1})\Phi(x,t_n)dx -$$

$$\int_{s_n}^{X} \frac{c_2\rho_2}{\Delta t} (u_n - u_{n-1})\Phi(x,t_n)dx$$

The last integrals can be rewritten as follows

$$\int_o^{s_n} \frac{(u_n - u_{n-1})}{\Delta t} \Phi(x,t_n)dx = -\int_o^{s_n} U_n \frac{\Phi(x,t_{n+1}) - \Phi(x,t_n)}{\Delta t} dx +$$

$$\int_o^{s_n} U_n \frac{\Phi(x,t_{n+1})}{\Delta t} dx - \int_o^{s_n} U_{n-1} \frac{\Phi(x,t_n)}{\Delta t} dx;$$

a similar expression holds over (s_n,X). If these expressions are substituted in the above equation, and the resulting expression is summed over n, we obtain

$$0 = \sum_{n=1}^{N} \Delta t\{\int_o^{s_n} U_N(L_1^*\Phi_N)(x,t_n)dx + \int_{s_n}^{X} U_N L_2^*\Phi_N(x,t_n)dx -$$

$$k_1\alpha(t_n)\Phi_x(0,t_n) + k_2 U_X \Phi_x(X,t_n)\} -$$

$$c_1\rho_1 \sum_{n=1}^{N} \{\int_o^{s_n} \frac{u_n\Phi(x,t_{n+1})}{\Delta t} dx - \int_o^{s_n} \frac{u_{n-1}\Phi(x,t_n)}{\Delta t} dx$$

$$-c_2\rho_2 \sum_{n=1}^{N} \{\int_{s_n}^{X} \frac{u_n\Phi(x,t_{n+1})}{\Delta t} dx - \int_{s_n}^{X} u_{n-1}\Phi(x,t_n)dx \equiv \hat{H}_N(U_N,\Phi_N,S_N)$$

The last two sums do not exactly telescope since the range of integration is changing; however, we observe that

$$\int_0^{s_n} u_n\Phi(x,t_{n+1})dx - \int_0^{s_{n+1}} u_n\Phi(x,t_{n+1})dx =$$

$$\int_{s_n}^{s_{n+1}} u_n(x)\Phi(x,t_{n+1})\,dx \leqq \max_{x\varepsilon(s_n,s_{n+1})}|u_n(x)|\,K\Delta t.$$

Moreover, since $u_n'' < \infty$ and $u_n'(s_n) < \infty$ uniformly in n and N, it follows that $|u_n(x)| \leq K\Delta t$ for $x \in (s_n,s_{n+1})$ for some constant K. Hence we see that

$$0 = H_N(U_N,\Phi_N,S_N) \to H_N(U_N,\Phi_N,S_N) \to H(U_N,\Phi_N,S_N) \to H(u,\Phi,s), \text{ or that}$$

$\{U_N,S_N\}$ converge to a weak solution of the Stefan problem.

Finally, by hypothesis there exists a unique weak solution of the Stefan problem. This implies that the originally computed sequence $\{U_N,S_N\}$, rather than the subsequence indexed by N_e and N_k, must converge to this solution. In other words, the numerical approximate solution converges to the weak solution of the Stefan problem.

Suggested exercises:

1) Verify all the convergence properties claimed in this section.

2) Prove that convergence of $\{U_{N_k},S_{N_k}\}$ and existence of a unique weak solution imply convergence of $\{U_N,S_N\}$ to this weak solution.